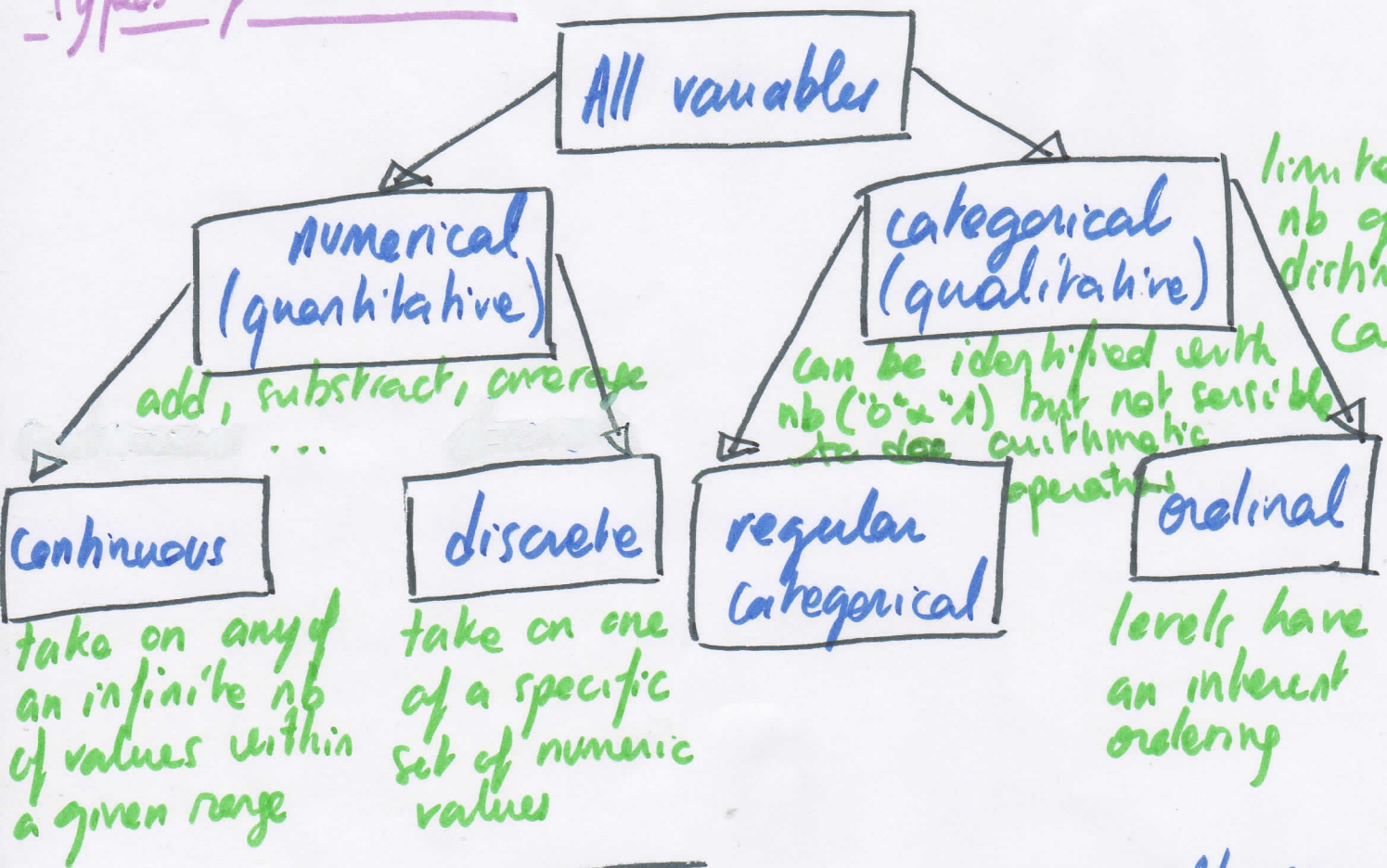
① DATA ANALYSIS & STATISTICAL INF - Duke

Types of variables



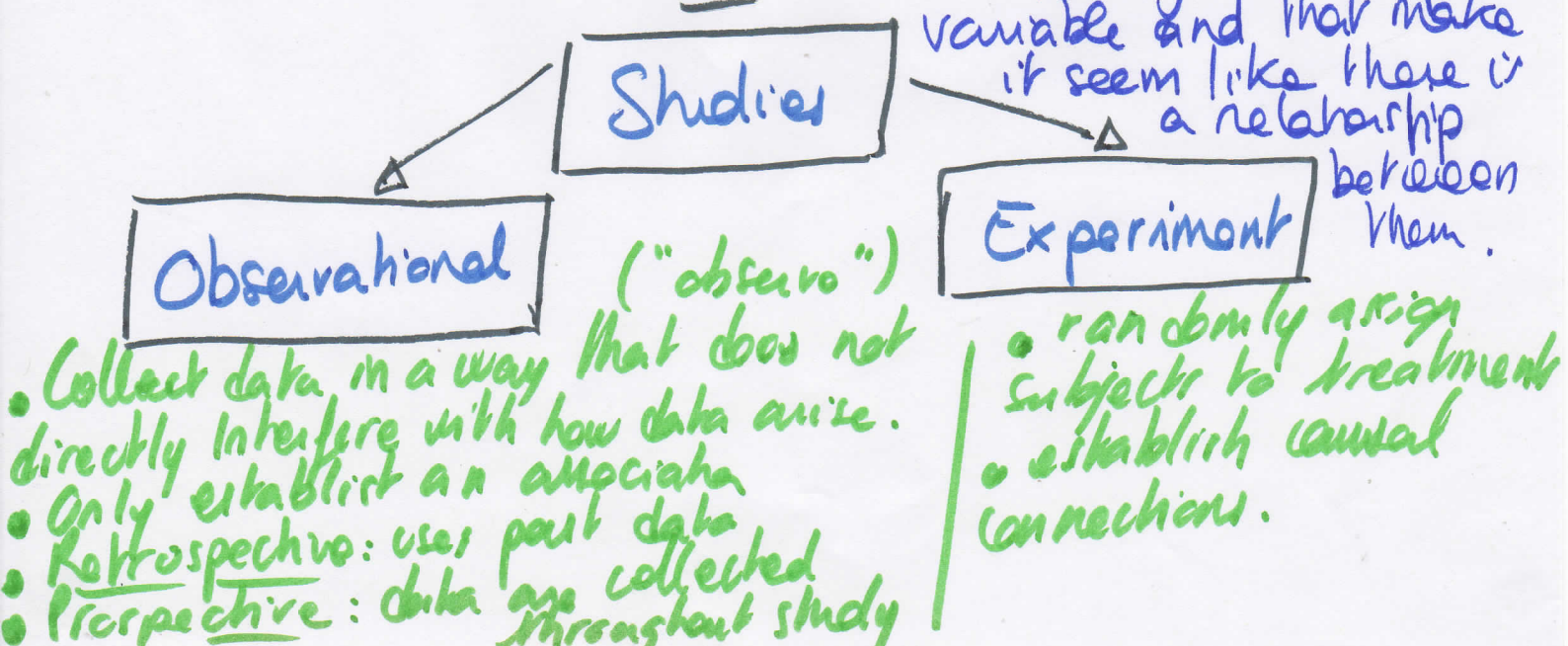
---

dependent variables: show some connection with one another (also called associated). Association ⊕ or ⊖

independent variables: not associated

Confounding variable: extraneous variables that affect both the explanatory & the response variable and that make it seem like there is a relationship between them.

Correlation does not imply causation



② DATA ANALYSIS & STATISTICAL INF - Duke

Census (recensement)

- Some individuals are hard to locate or measure, and these people be different from the rest of the pop.
- Populations rarely stand still

Sample bias

- Convenience sample: individuals who are easily accessible are more likely to be included in the sample.
- Non-response: If only a (non-random) fraction of the randomly sampled people respond to a survey then that the sample is no longer representative of the pop.
- Voluntary response: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue.

Sampling methods

Simple Random Sample (SRS): each case is equally likely to be selected

Stratified sample: Divide the population into homogenous strata, then randomly sample from within each stratum

Cluster sample: Divide the population into clusters, randomly sample a few clusters, then randomly sample from within those clusters. (eg. geographical)

⚠ Clusters are not really homogenous, contrary of stratum

If you generalize an analysis to the entire population this is an inference making

* + Multistage sample cluster + sample within each cluster *

⚠ Observational studies can allow us to infer correlation & experimental studies causation

③ DATA ANALYSIS & STATISTICAL INF Duke

Principle of experimental design

① Control: Compare treatment of interest to a control group.
② Randomize: Randomly assign subjects to treatments
③ Replicate: collect a sufficiently large sample, or replicate the entire study
④ Block: block for variables known or suspected to affect the outcome.

Blocking vs. explanatory variable

Explanatory variables (factors): conditions we can impose on experimental units
Blocking variables: characteristics that the experimental units come with, that we would like to control for

- Blocking during random assignment is like stratifying during random sampling

Experimental terminology

Fake treatment, often used as the control group for medical studies = Placebo
Placebo effect: showing change despite being on the placebo
Blinding = Experimental units don't know which group they are in
Double blind: both the experimental units and researchers don't know the group assignment.

(4) DATA ANALYSIS & STATISTICAL INF. - Duke

Random sampling vs. assignment

Random sampling: (1)

generalizability →

sample

Random assignment: (2)

Causality

Most observational studies

ideal experiment but difficult to conduct

	Random assignment	No Random assign.	
Random Sampling	Causal & generalizable	not causal, but generalizable	Generalizability
No Random Sampling	Causal, but not generalizable	Neither causal nor generalizable	No generalizability
	Causation	Association	

Most experiments

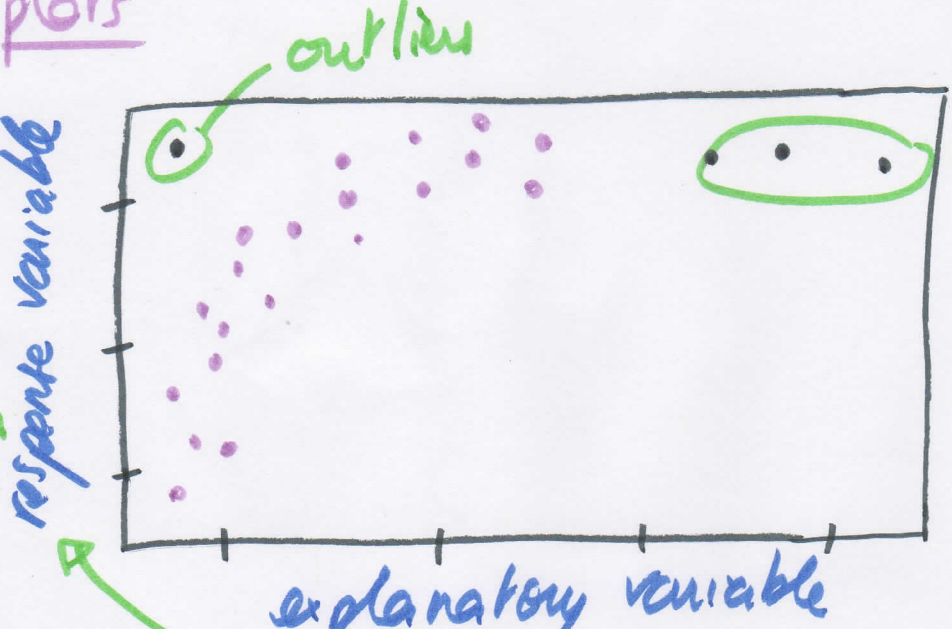
bad observational studies

Visualizing numerical data

- Scatterplots for paired data
- other visualizations for describing distributions of numerical variables

⑤ DATA ANALYSIS & STATISTICAL INF Duke

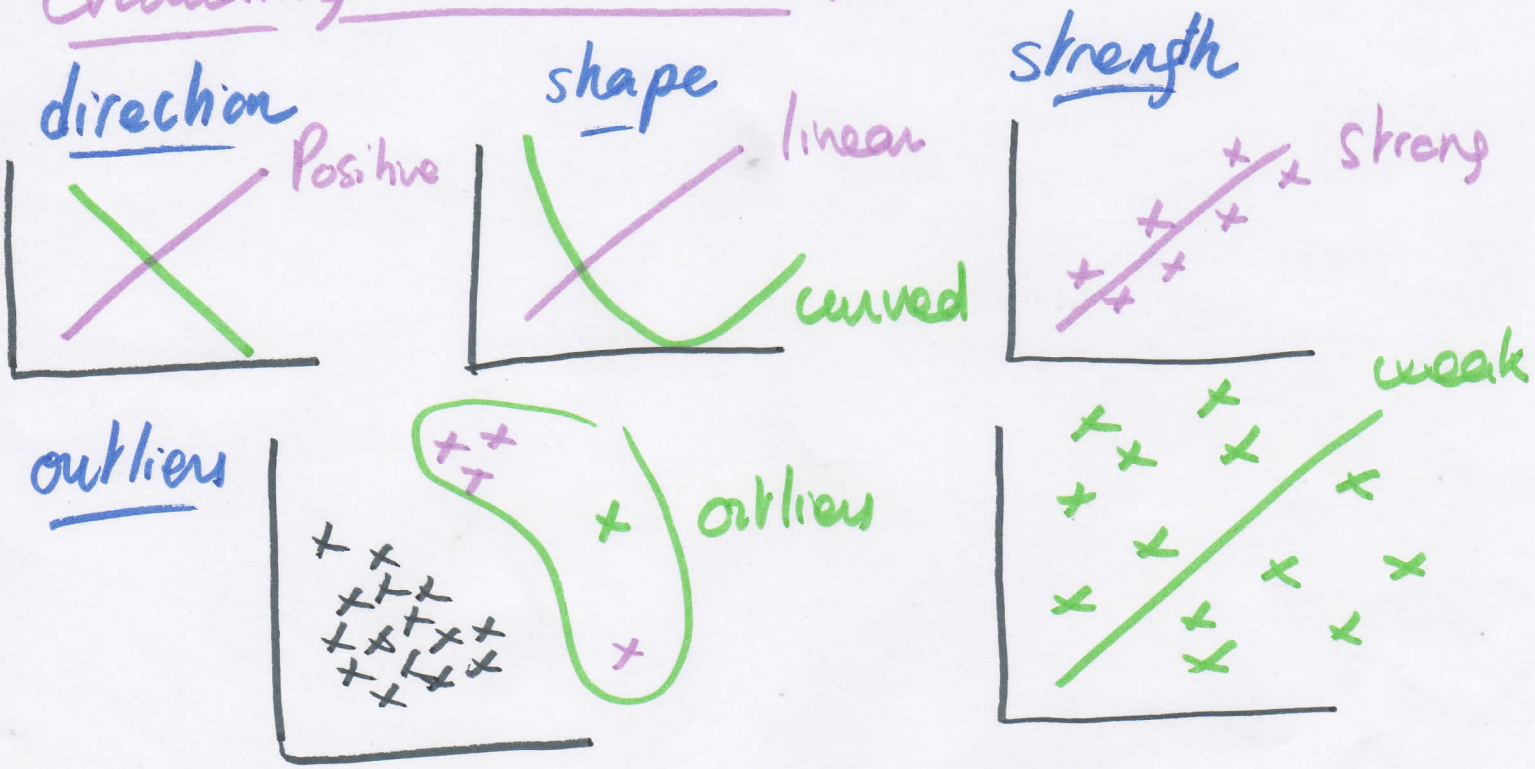
Scatterplots → two numerical variables



correlation
not causation

the relationship

Evaluating the relationship



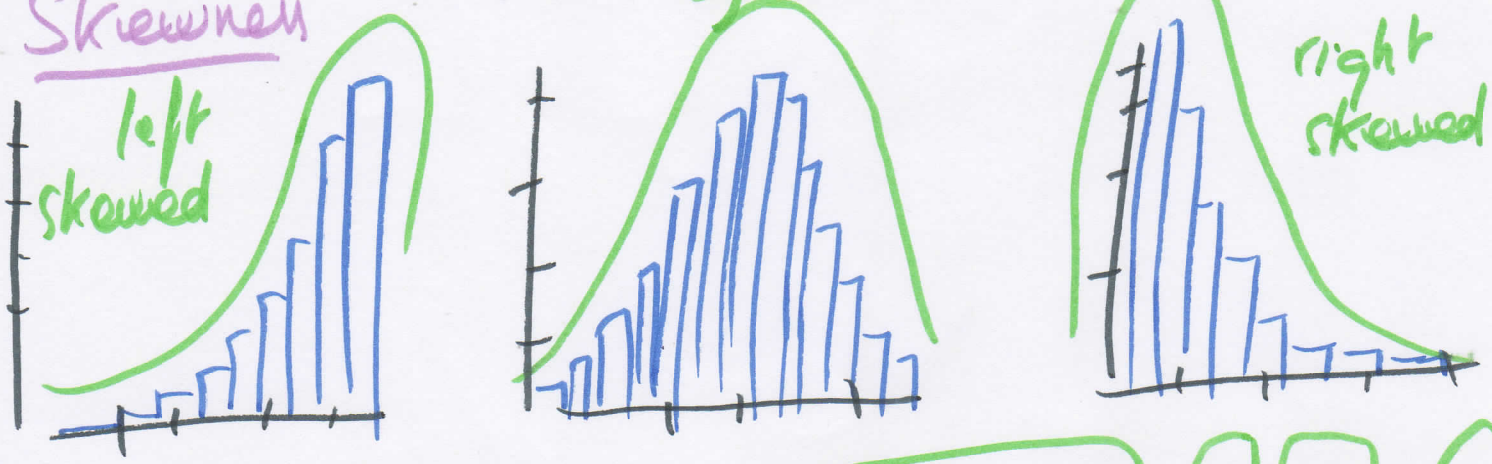
Histogram

- Provide a view of the data density
- Especially useful for describing the shape of the distribution.

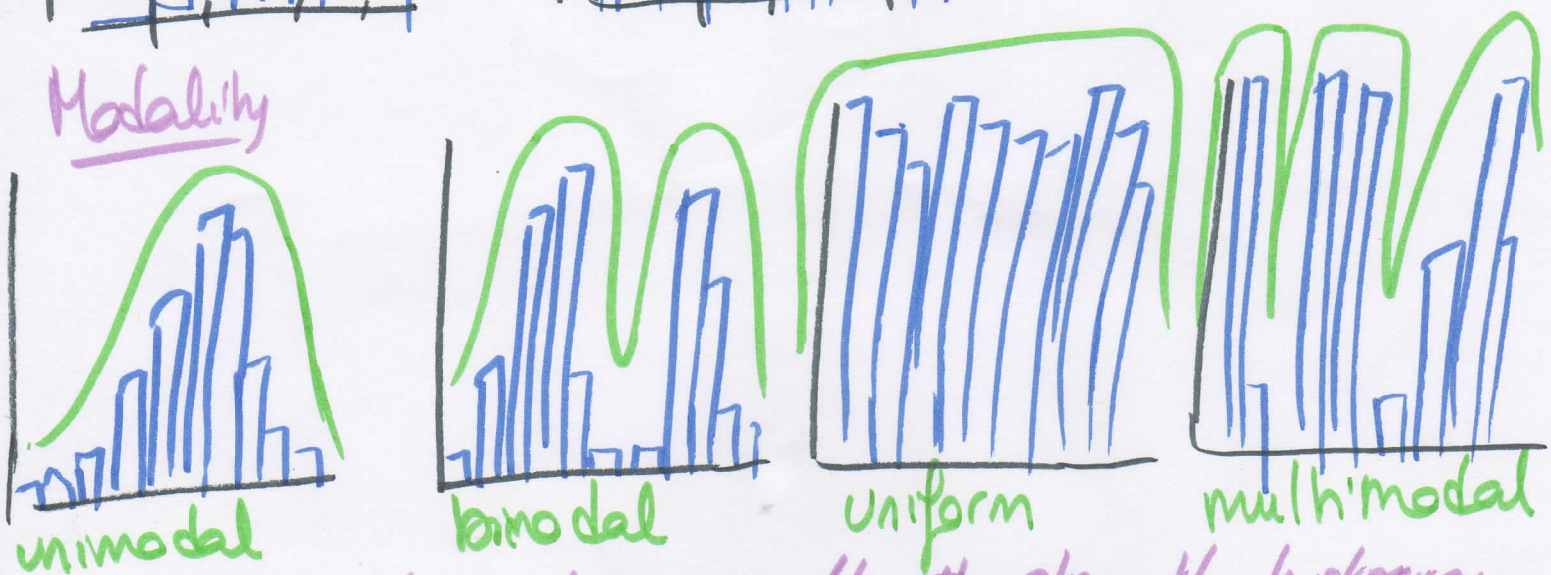
DATA ANALYSIS & STATISTICAL INF Duke

(6)

Skewness



Modality



⇒ The chosen bin width can alter the story the histogram is telling

Dotplot

- Useful when individual values are of interest
- Can get busy as the sample size increases

Boxplot

- Useful for highlighting outliers, media, IQR
- Can show the skewness of data as histograms

② DATA ANALYSIS & STATISTICAL INF Duke

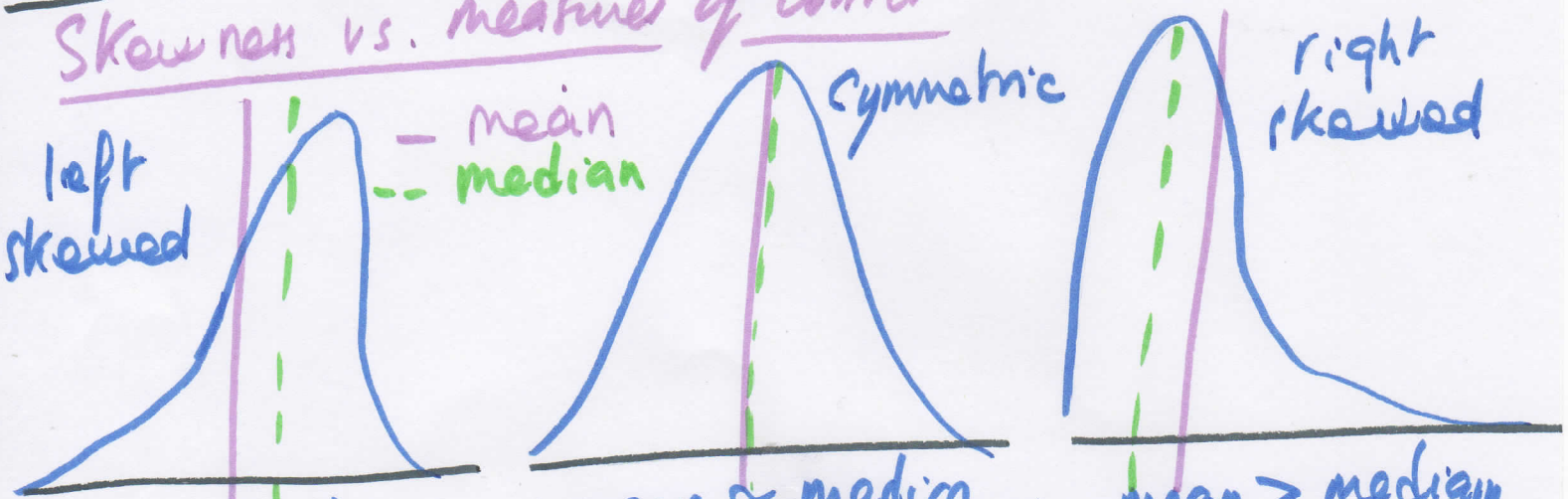
Measures of center

Mean: arithmetic average — $\bar{x}$ sample mean, μ population mean

Median: midpoint of the distribution (50th percentile)

Mode: Most frequent observation

Skewness vs. measures of center



affected by outliers

Measures of spread

① Range (max - min)

② Variance: $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ ② get rid of ⊖ and increase larger deviation

Squared deviation from the mean

s^2 = Sample variance

σ^2 = population variance

③ Standard deviation: $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

s or σ

same units as the data

av. deviation around the mean

• variability ≠ diversity

better than range because not affected by outliers or unusual values (see box plot)

④ Interquartile range (IQR)

Range of the middle 50% of the data, distance between the first quartile (25th percentile) & third quartile (75th percentile)

⑧ DATA ANALYSIS & STATISTICAC INF Duke

Robust statistics

	robust	non robust
Center	median	mean
Spread	IQR	standard deviation range

useful for symmetric distributions

useful for skewed, with extreme observations

Transforming data

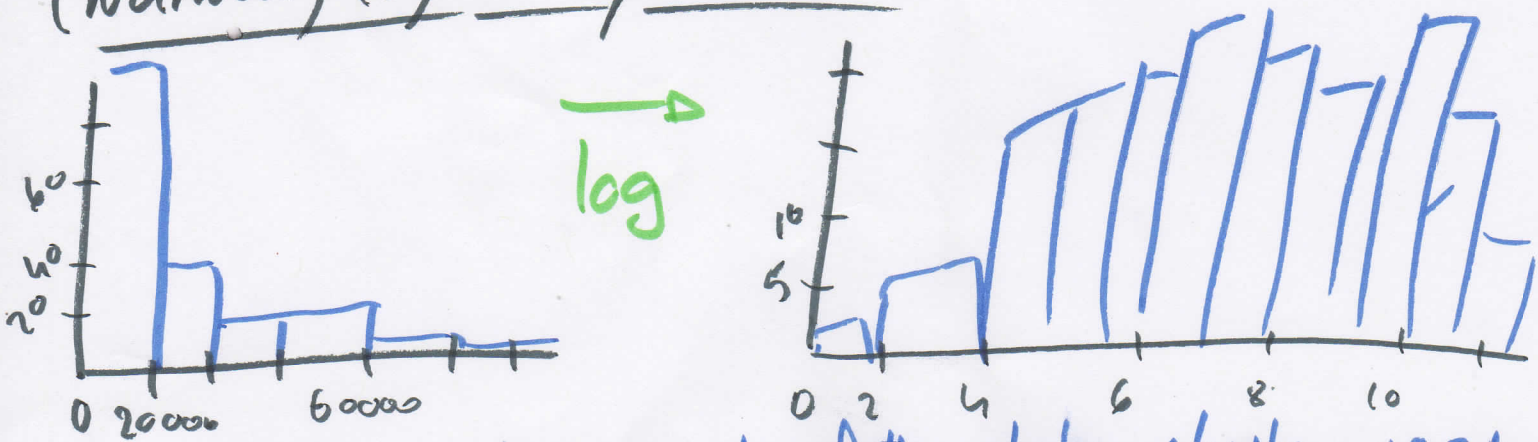
Goals: . see the data structure differently
. Reduce skew assist in modeling
. Straighten a non-linear relationship in a scatterplot

⇒ A transformation is a rescaling of the data using a function.

When data are very strongly skewed, we sometimes transform them so they are easier to model.

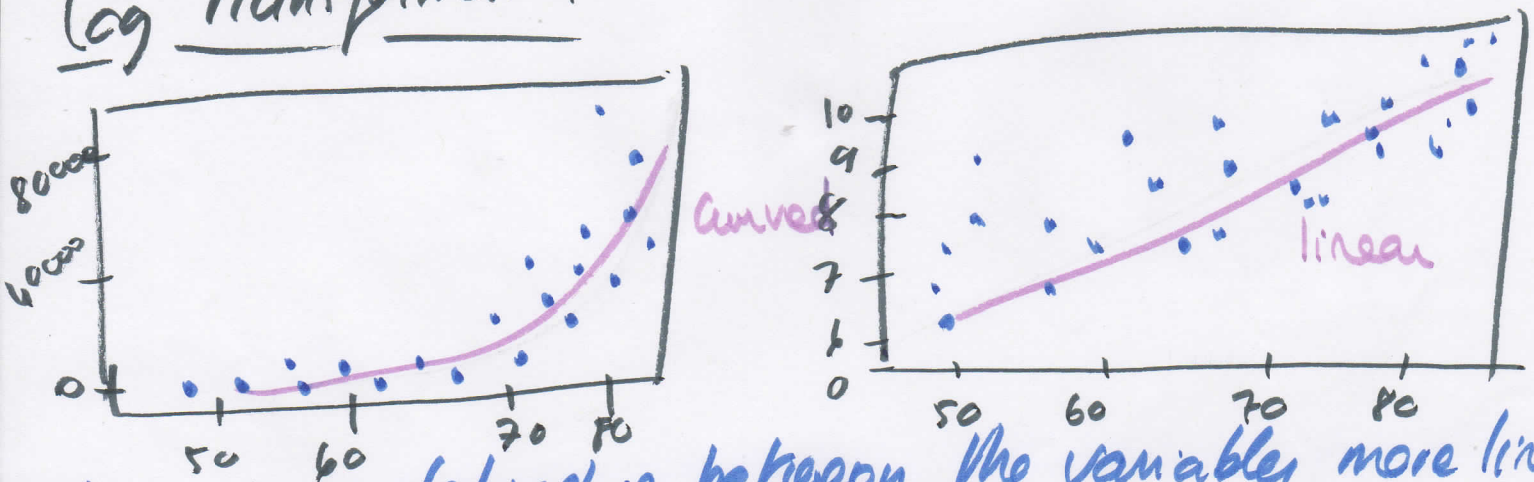
⑨ DATA ANALYSIS & STATISTICAL INF - Duke

(Natural) log transformation



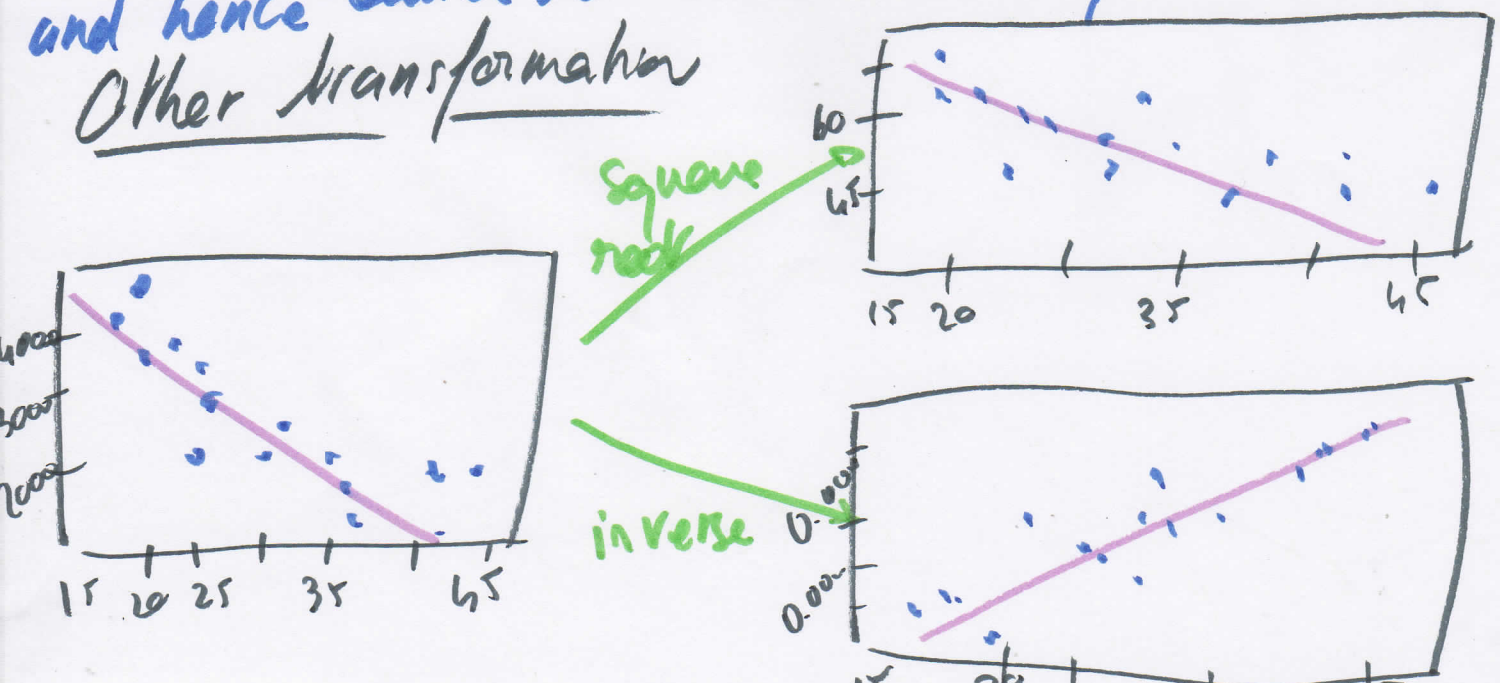
often applied when much of the data clusters near zero (relative to the larger values in the data set) and all observations are positive.

log transformation



Make the relationship between the variables more linear and hence easier to model with simple methods.

Other transformation



⑩ DATA ANALYSIS & STATISTICAL INF Duke

Exploring categorical variables

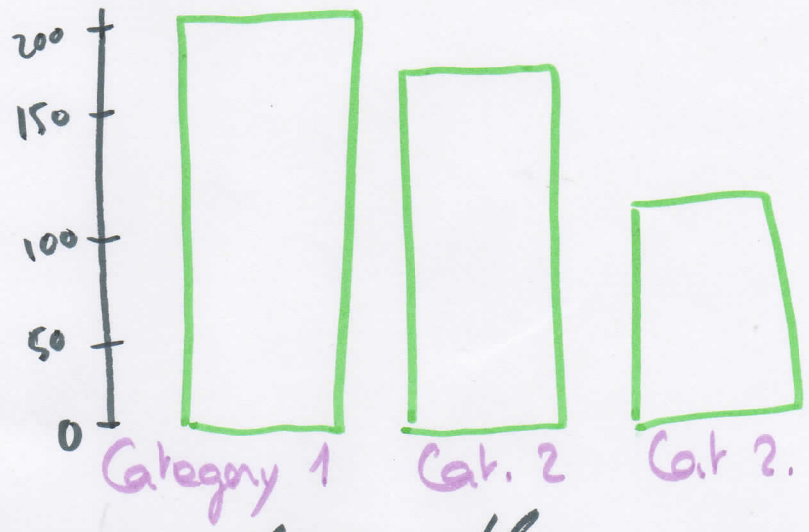
① Describe distribution of a single categorical variable

- no histogram
- no pie chart

Frequency table

Categorical variable	Counts	Frequencies (%)
TOTAL	number	100 %

Bar plot



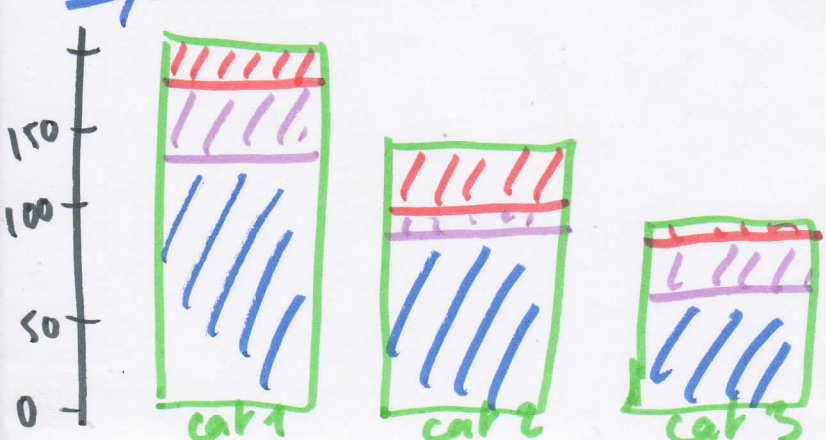
② Evaluate relationship between two categorical variables

Contingency table

Categorical var 1	categorical	var. 2	TOTAL
	nb	nb	nb
	...	...	nb
	...	...	nb
TOTAL	nb	nb	nb Tot

← calculate relative frequencies nb/total

Segmented bar plot

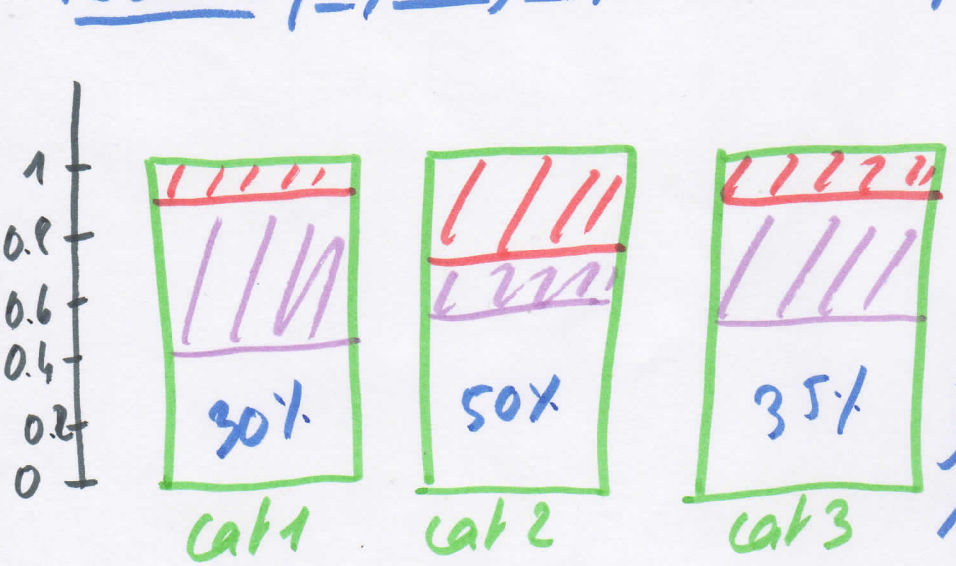


- useful for visualizing conditional frequency distribution but not compare the relative freq.

DATA ANALYSIS & STATISTICAL INF Duke

(11)

Relative frequency segmented bar plot

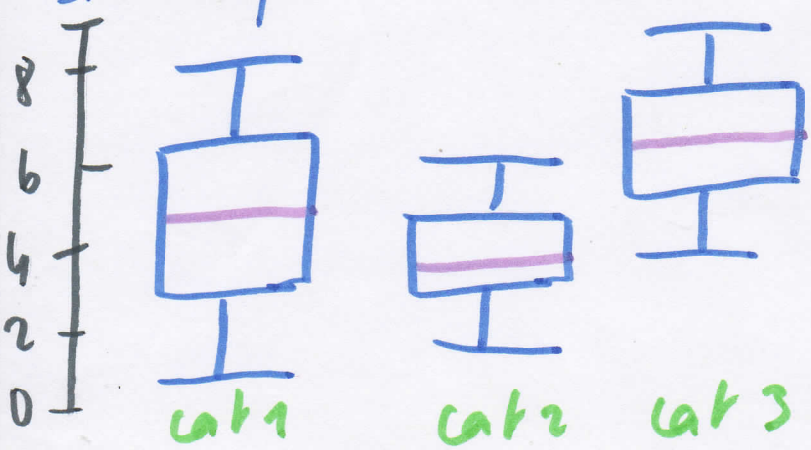


- Compare relative frequencies to explore the relationship between the variables.

Mosaicplot

(3) Evaluate relationship between a categorical and a numerical variable

Side-by-Side box plots



[look at (19) for next "Intro to Inference"]

(12) DATA ANALYSIS & STATISTICS

Dan

Random process: we know what outcomes could happen, but we don't know which particular outcome will happen.

Probability $P(A)$ = probability of event A

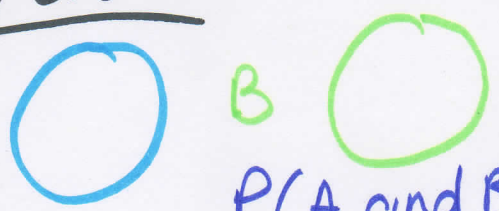
$0 \leq P(A) \leq 1$

- **frequentist interpretation**: the probability of an outcome is the proportion of time the outcome would occur if we observed the random process an infinite number of times.
- **bayesian interpretation**: interprets probability as a subjective degree of belief.

Largely popularized by revolutionary advance in computational technology & methods during the last twenty years

law of large numbers: as more observations are collected the proportion of occurences with a particular outcome converges to the probability of that outcome.

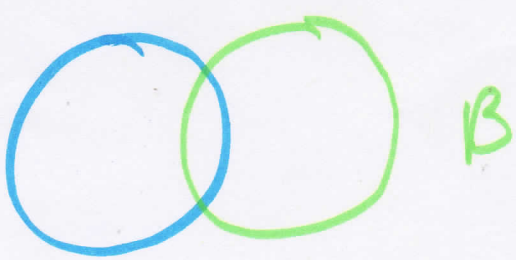
disjoint (mutually exclusive) events: cannot happen at the same time



$P(A \text{ and } B) = 0$

non-disjoint events:
can happen at the same time

$P(A \text{ and } B) \neq 0$



(13) DATA ANALYSIS & STATISTICAL INF — Duke

Union of: disjoint events — $P(A \text{ or } B) = P(A) + P(B)$

non disjoint events — general addition rule: $P(A) + P(B) - P(A \text{ and } B)$

Sample space: collection of all possible outcomes of a trial. E.g. with Male & Female $S = \{MM, FF, FM, MF\}$

Probability distributions: lists all possible outcomes in the sample space, and the probabilities with which they occur.

One toss	head	tail
prob.	0.5	0.5

L. Rules:
1. The events listed must be disjoint
2. Each probability must be between 0 and 1
3. The probability must total 1

Complementary events: two mutually exclusive events whose probabilities add up to 1. + (1a)

Independance: two processes are independent if knowing the outcome of one provides no useful information about the outcome of the other.

$P(A|B) = P(A)$, then A & B are independent

($P(A|B)$: given B occured)

$P(A) \times P(B) = P(A \text{ and } B)$

Conditional probability: $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

(Bayes' theorem) Hen $P(A \text{ and } B) = P(A|B) \times P(B)$

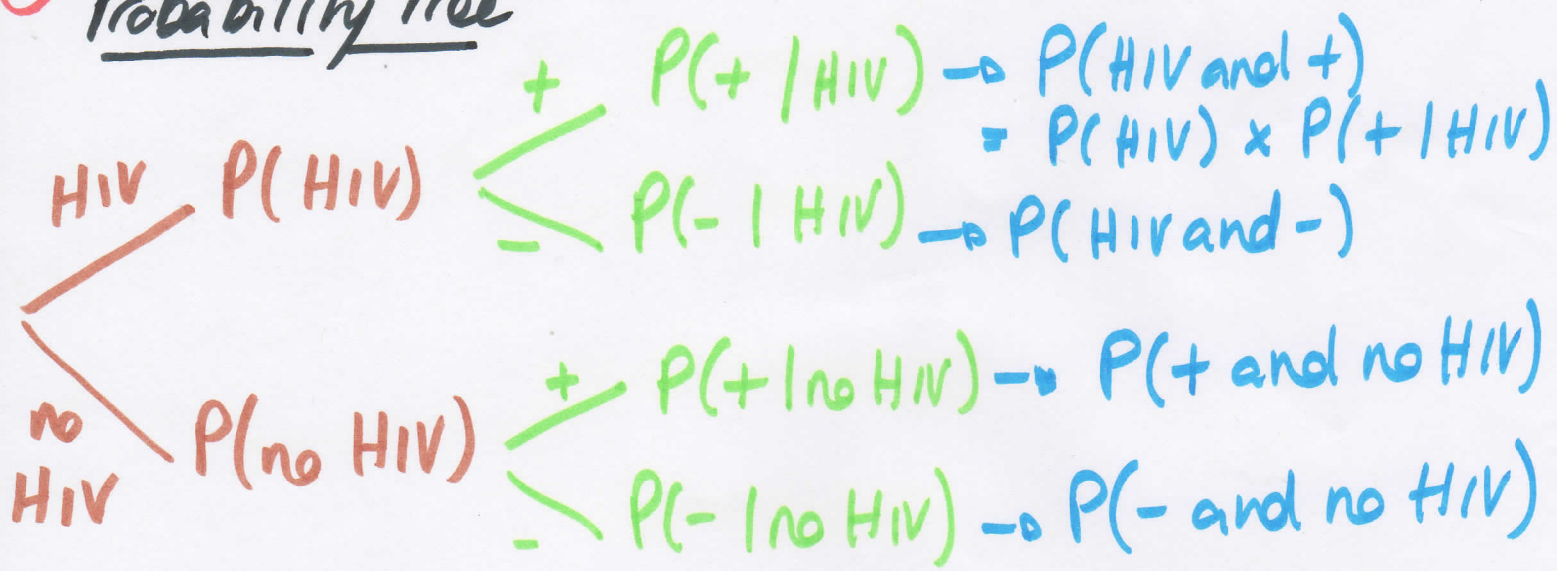
If $P(A|B) = P(A)$ then the events A and B are said to be independent.

Ind: $P(A \cup B) = P(A) + P(B) - P(A) \times P(B)$

DATA ANALYSIS & STATISTICAL INF

Duke

(14) **Probability tree**



Bayesian inference

1. Set a prior
2. Collect data
3. Obtain a posterior
4. Update the prior with the previous posterior

Posterior: Probability we set forth, given the data we just observed. P(hypothesis | data)

Standardized (Z) score: of an observation is the number of standard deviations it falls above or below the mean. Defined for distributions of any shape

Z score of mean = 0 unusual observation: $|Z| > 2$

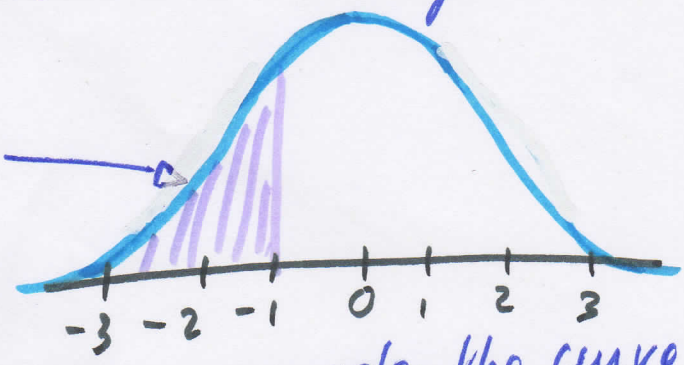
$$Z = \frac{\text{observation} - \text{mean}}{SD}$$

Percentiles: When the distribution is normal, Z scores can be used to calculate percentiles.

DATA ANALYSIS & STATISTICAL INF

(15) Duke

Percentile: percentage of observations that fall below a given data point.
Graphically, percentile is the area below the probability distribution curve to the left of that observation



E.g: ① mean: 1500 sd: 300 for probability / cutoff value observation: 1800 area under the curve
What is the percentile score of that observation?
In R: pnorm(1800, mean = 1500, sd = 300) : 0,8413 for normal distribution
Or use a probability table

② We are looking for an unknown observation.
mean: 1500 percentile score: 90% sd: 300
We look for the standardized (Z) score on a probability table (looking for about 90% p. score) and calculate $Z = 1.28 = \frac{X - 1500}{300}$ $X = 1884$

Normal distribution: unimodal and symmetric. bell curve
- follows very strict guidelines about how variably the data are distributed around the mean.
- Many variables are nearly normal, but none are exactly normal. 2 parameters: $N(\mu, \sigma)$ mean, standard deviation

68-95-99.7 % rule

Anatomy of a normal probability plot

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a one-to-one relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution.
- Since a one-to-one relationship would appear as a straight line on a scatter plot, the closer the points are to a perfect straight line, the more confident we can be that the data follow the normal model.
- Constructing a normal probability plot requires calculating percentile and corresponding z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.

Short tails: narrower than the normal distribution. Points follow an S-shaped-curve

Long tails wider than the normal distribution. Points start below the line, bend to follow it and end above it.

DATA ANALYSIS & STATISTICAL INF.

Duke

(17) Finding the cut off value / observation of a normal distribution

mean: 77 sd: 5 percentile: 0.20

In R: qnorm(0.20, mean = 77, sd = 5) — qnorm for quantile or cut off value

---

Binomial distribution: describes the probability of having exactly k successes in n independent Bernoulli trials with probability of success p.

"n choose k"

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

In R: choose(n, k)

# of scenarios $\times$ P(single scenario)

$p^k (1-p)^{(n-k)}$

p^k: probability of success to the power of number of successes

$(1-p)^{(n-k)}$: probability of failure to the power of number of failure

Examples:

How many scenarios yield 1 success in 4 trials? $n = 4$ $k = 1$

$$\binom{4}{1} = \frac{4!}{1! \times (4-1)!} = \frac{4 \times 3 \times 2 \times 1}{1 \times 3 \times 2 \times 1} = 4$$

How many scenarios yield 2 successes in 9 trials?

$n = 9$ $k = 2$ $\binom{9}{2} = \frac{9!}{2!\,7!} = \frac{9 \times 8 \times 7!}{2 \times 1 \times 7!} = 36$

Binomial distribution: If p represents probability of success, $(1-p)$ represents probability of failure, n represent number of independent trials, and k represents nb of successes

(18) DATA ANALYSIS & STATISTICAL INF Duke

$P(K \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{(n-k)}$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Binomial conditions

1- The trials must be independent
2- The number of trials, n, must be fixed
3- each trial outcome must be classified as a success or a failure
4- The probability of success, p, must be the same for each trial

Example if $n = 20$ $k = 5$ $p = 0.262$

Probability: $P(k=5) = \binom{20}{5} 0.262^5 \times 0.738^{15}$

In R: `dbinom(k, size = 20, p = 0.262)` → can be a:b for $P(X \geq a)$. use sum()

Expected value (mean) of binomial distribution:

$$\mu = np$$

Standard deviation of binomial distribution:

$$\sigma = \sqrt{np(1-p)}$$

Success-failure rule: A binomial distribution with at least 10 expected successes and 10 expected failures closely follows a normal distribution $np \geq 10$ $n(1-p) \geq 10$

Normal approximation to the binomial: If the success-failure condition holds, $\text{Binomial}(n, p) \sim \text{Normal}(\mu, \sigma)$ where $\mu = np$ and $\sigma = \sqrt{np(1-p)}$

DATA ANALYSIS & STATISTICAL INF

(19) (11) Duke

[part next to (11)] Introduction to Inference

① **Null hypothesis**: "there is nothing going on" H_0
Alternative hypothesis: "there is something going on" H_A
② **Present the evidence**: collect data
③ **Judge the evidence**: "Could these data plausibly have happened by chance if the null hypothesis were true"?
If yes: fail to reject H_0 **If no**: Reject H_0

Hypothesis testing framework

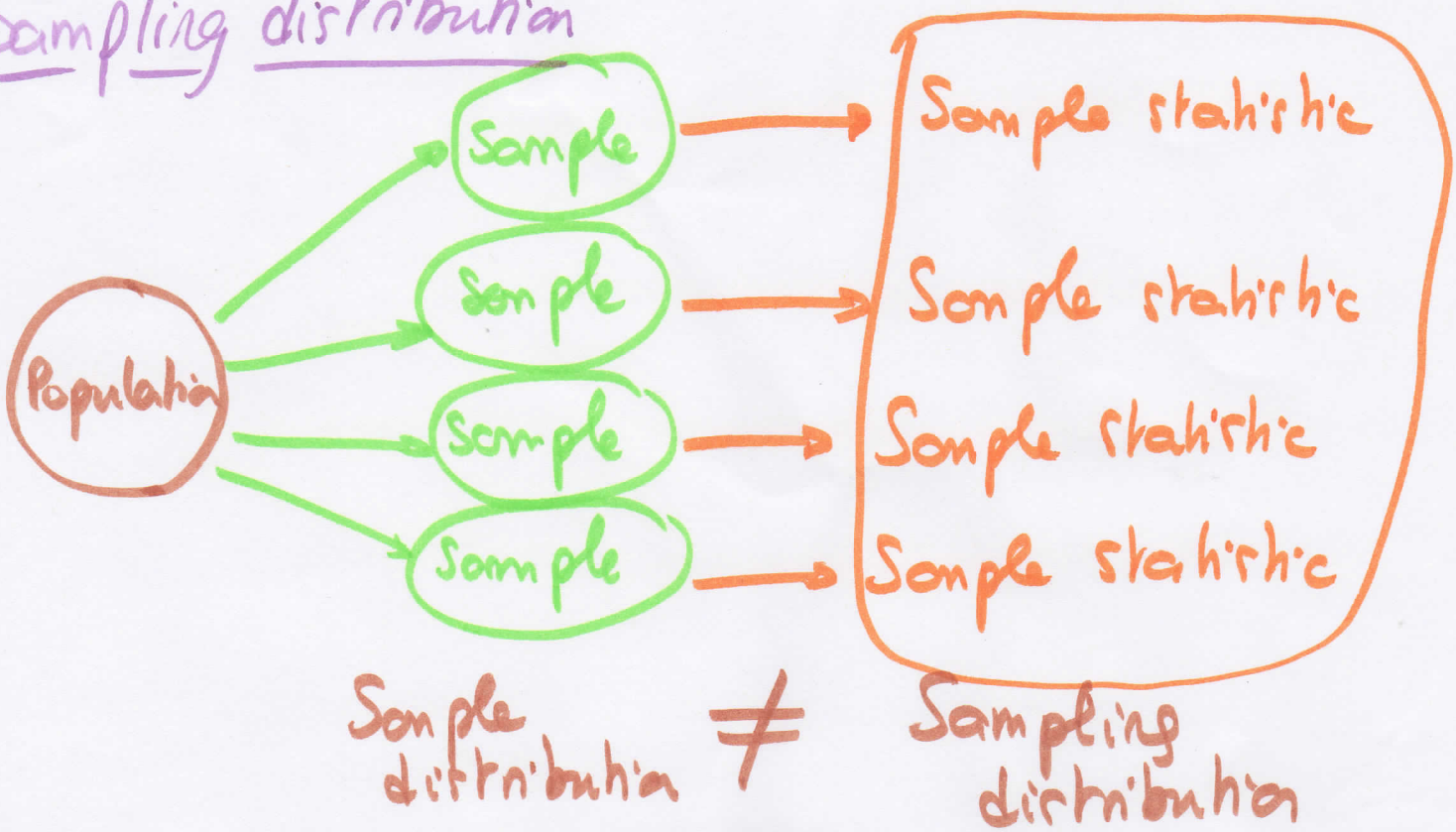
① Start with a **null hypothesis** (H_0) that represents the status quo
② Set an **alternative hypothesis** (H_A) that represents the research question, i.e. what we are testing for
③ Conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or theoretical methods
- if the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, stick with the null hypothesis → independent variables
- if they do, then reject the null hypothesis in favor of the alternative → dependent variable / relation

④ **Complementary** events are always necessarily **disjoint** by definition. Disjoints events aren't always complementary

(20) DATA ANALYSIS AND STATISTICAL INFERENCE

Duke

Sampling distribution



Sample distribution ≠ Sampling distribution

Central Limit Theorem (CLT): The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

standard error

$$\bar{x} \sim N(\text{mean} = \mu, \; SE = \frac{\sigma}{\sqrt{n}})$$

shape ↓ center ↓ spread ↓

σ: sample sd because population sd almost impossible

Conditions for the CLT:

① **Independence**: Sampled observations must be independent.
- random sample / assignment
- if sampling without replacement, $n < 10\%$ of population

② **Sample size / skew**: either the population distribution is normal, or if the population distribution is skewed, the

(21) sample size is large.

Confidence interval for a population mean (CI): Computed as the sample mean plus/minus a margin of error (critical value corresponding to the middle XX% of the normal distribution times the standard error of the sampling distribution).

point estimate (here sample mean) $\bar{x}$ $\pm$ critical value z^* $\frac{s}{\sqrt{n}}$ standard error

margin of error

Conditions for this confidence interval:

① Independence: sampled observations must be independent
- random sample/assignment
- if sampling without replacement, $n < 10\%$ of population

② Sample size/skew: $n \geq 30$, larger if the population distribution is very skewed → large sample

---

Backtracking to n for a given margin of error:

Given a target margin of error, confidence level, and information on the variability of the sample (or the population), we can determine the required sample size to achieve the desired margin of error.

$$ME = z^* \frac{s}{\sqrt{n}} \Rightarrow \text{required sample size } n = \left(\frac{z^* s}{ME}\right)^2$$

ME: margin of error

if we want to decrease the margin by 2 (/2):

$$\frac{1}{2} ME = z^* \frac{s}{\sqrt{n}} * \frac{1}{2} \Rightarrow \frac{1}{2} ME = z^* \frac{s}{\sqrt{4n}}$$

DATA ANALYSIS & STATISTICAL INFERENCE

Duke

(22)

Sampling distribution

Sample means: mean of $\bar{x}$ → sample mean

Standard deviation of sample means = standard error → sd of $\bar{x}$

The higher the sample size of the samples, the lower the variation of the sample means.

Central Limit Theorem

The more the skew, the higher the sample size is needed for the Central Limit Theorem to kick in. ($n \geq 30$)

Confidence interval: a plausible range of values for the population parameter.

- if we report a point estimate, we probably won't hit the exact population parameter.
- if we report a range of plausible values we have a good shot at capturing the parameter.

Finding the critical value 95% confidence

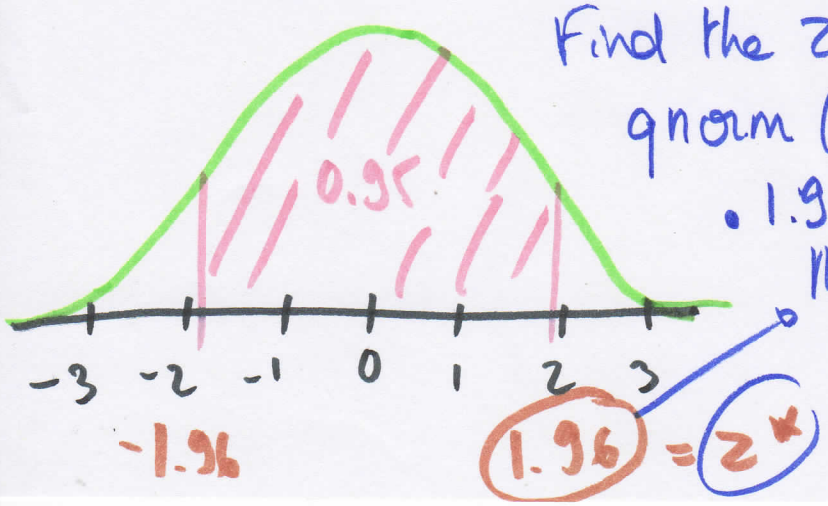
$$\bar{x} \pm z^* \frac{s}{\sqrt{n}}$$

Find cutoff value = $(1 - 0.95)/2$ ← total under the curve

$= 0.025$

Find the z score value of 0.025

qnorm(0.025) → -1.96



- 1.96 is the positive cutoff. for the middle 95% of the standard normal distribution.

Accuracy: whether or not the CI contains the true population parameter.

Precision: width of a CI

- Commonly used confidence levels in practice are 90%, 95%, 98% and 99%.
- Changing the confidence level simply means adjusting the value of the critical value in the CI formula.
- The higher the confidence level, the larger the critical value, hence the larger the margin of error, and hence the width of the confidence interval and the accuracy
- If we want to increase the accuracy, we need to increase the confidence level. But it can come at a cost, precision
- CL ↗ width ↗ accuracy ↗ precision ↘
- How can we get the best of both world - higher precision and higher accuracy?
  → increase sample size
- All else held constant, as sample size increases, the margin of error decreases.

Confidence interval examples

Based on responses from 1151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.

Interpretation: We are 95% confident that Americans on average have 3.40 to 4.24 bad mental health days per month.

In this context, what does a 95% confidence level mean?
95% of random samples of 1151 American will yield

(24) → CIs that capture the true population mean of number of bad mental health days per month.

Hypothesis testing (for a mean):

Null - H_0 often either a skeptical perspective or a claim to be tested. = value → represents the research Q

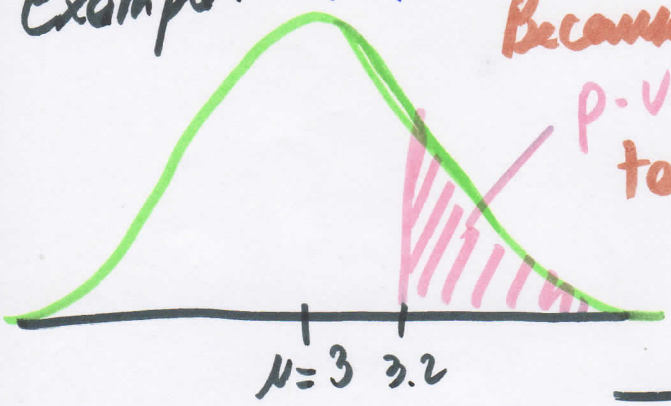
Alternative - H_A Represents an alternative claim under consideration and is often represented by a range of possible parameter values $<, >, \neq$ value

- The skeptic will not abandon the H_0 unless the evidence in favor of the H_A is so strong that she rejects H_0 in favor of H_A.
- The hypothesis are always about population parameters, never about sample statistics. (because we don't know the pop. parameter)

p-value: P(observed or more extreme outcome | H_0 true)

Example: $P(\bar{X} > 3.2 \mid H_0: \mu = 3)$ $\bar{X} \sim N(\mu = 3, SE = 0.246)$

Because we assume H_0 true



p-value

test statistic: $Z = \frac{3.2 - 3}{0.246} = 0.81$

p-value = $P(Z > 0.81) = 0.209$

Since p-value is high, we do not reject H_0 * (25)

Decision based on the p-value

- We used the test statistic to calculate the p-value, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.

DATA ANALYSIS & STATISTICAL INFERENCE

Duke

(23)

- If the p-value is low (lower than the significance level α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true and hence reject H_0.
- If the p-value is high (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence do not reject H_0.

---

(24) Interpreting the p-value ($\simeq 0,21$)

- If in fact college students have been in 3 exclusive relationships on average, there is a 21% chance that a random sample of 50 college students would yield a sample mean of 3.2 or higher.
- This is a pretty high probability, so we think that a sample mean of 3.2 or more exclusive relationships is likely to happen simply by chance.

Making a decision

- Since p-value is high (> 5%) we fail to reject H_0.
- These data do not provide convincing evidence that college students have been in more than 3 relationships on average.
- The difference between the null value of 3 relationships and the observed sample mean of 3.2 relationships is due to chance or sampling variability.

Two sided tests

- Often instead of looking for a divergence from the null in a specific direction, we might be interested in divergence in any direction.

. We call such hypothesis test two-sided (or two-tailed).

• The definition of a p-value is the same regardless of doing a one or two-sided test, however the calculation is slightly different since we need to consider "at least as extreme as the observed outcome" in both direction.

Example: $P(\bar{X} > 3.2 \text{ or } \bar{X} < 2.8 \mid H_0: \mu = 3)$

p-value = $P(Z > 0.81) + P(Z < -0.81)$
$= 2 \times 0.209$
$= 0.418$

Hypothesis testing for a single mean (summary)

1. Set the hypotheses: $H_0: \mu =$ null value
   $H_A: \mu <$ or $> \mu$ or $\neq$ null value
2. Calculate the point estimate: $\bar{x}$
3. Check conditions: independance / sample size & skew
4. Draw sampling distribution, shade p-value, calculate test statistic $Z = \frac{\bar{x} - \mu}{SE}$, $SE = \frac{s}{\sqrt{n}}$
5. Make a decision and interpret it in context of the research question:

• If p-value $< \alpha$, reject H_0; the data provide convincing evidence for H_A.
• If p-value $> \alpha$, fail to reject H_0 the data do not provide convincing evidence for H_A

DATA ANALYSIT & STATISTICAL INFERENCE Duke

(L7)

Unbiased estimator

- An important assumption about point estimates is that they are **unbiased**, i.e. the sampling distribution of the estimate is centered at the true population parameter it estimate.
- An unbiased estimate does not naturally over or underestimate the parameter, it provides a "good" estimate.
- Example: Sample mean

Decision error: Goal: Keep α and β low

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	Type 1 error, α
	H_A true	Type 2 error, β	$1-\beta$

- **Type 1 error** is rejecting H_0 when H_0 is true phab. α
- **Type 2 error** is failing to reject H_0 when H_A is true β
- We (almost) never know if H_0 or H_A is true, but we need to consider all posibilities.

Type 1 error rate

- We reject H_0 when the p-value is less than 0.05 ($\alpha = 0.05$)
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times. . In other words, when using a 5% significance level

DATA ANALYSIS & STATISTICAL INFERENCE

Duke

(22) → There is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of α.

Increasing α increases the Type 1 error rate.

Choosing α

- If Type 1 error is dangerous or especially costly, choose a small significance level (eg 0.01)

Goal: We want to be very cautious about rejecting H_0, so we demand very strong evidence favoring H_A before we would do so.

- If a type 2 error is relatively more dangerous or much more costly, choose a higher significance level (e.g. 0.10)

Goal: We want to be cautious about failing to reject H_0 when the null is actually false.

Goal in decision error

- Keep α and β low

α = probability of type 1 error

β = probability of type 2 error

Effect size (δ): difference between point estimate and null value. β depends on it. The null hypothesis is more likely to be rejected if the effect size is large.

(29)

Agreement of Confidence Interval and Hypothesis testing

- A two sided hypothesis with threshold of α is equivalent to a confidence interval with $CL = 1 - \alpha$
- A one sided hypothesis with threshold of α is equivalent to a confidence interval with $CL = 1 - (2 \times \alpha)$
- If Ho is rejected, a confidence interval that agree with the result of the hypothesis test should not include the null value
- If Ho is failed to be rejected, a confidence interval that agree with the result of the hypothesis test should include the null value.

Statistical vs. practical significance

- Real differences between the point estimate and null value are easier to detect with larger samples.
- However, very large sample will result in statistical significance even for tiny differences between the sample mean and the null value (effect size), even when the difference is not practically significant.
- In order to make sure that the findings do not suffer from the problem of being statistically significant but not practically significant, it is good to do an a priory analysis to decide what sample size to chose based on characteristics from the population.

The standard error measures the variability of sample statistics (the variability in point estimates from different samples of the same size and from the same population).

DATA ANALYSIS & STATISTICAC INFERENCE Duke

(30) Central Limit Theorem for the mean (cumary)

(1a) The sample size is sufficiently large ($n \geq 30$) and the data are not extremely skewed or

(1b) The population is known to have a normal distribution and

(2) The observations in the sample are independent, then the distribution of the sample mean will be nearly normal, centered at the true population mean and with a standard error of $\frac{\sigma}{\sqrt{n}}$

$$\bar{x} \sim N\left(\text{mean} = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

- When the population distribution is unknown, condition (1a) can be checked using a histogram or some other visualization of the distribution of the observed data in the sample.
- The larger the sample size (n), the less important the shape of the distribution becomes, i.e. when n is very large the sampling distribution will be nearly normal regardless of the shape of the population distribution.

Interpret a confidence interval

- "We are XX % confident that the true population parameter is in this interval" where XX% is the desired confidence level

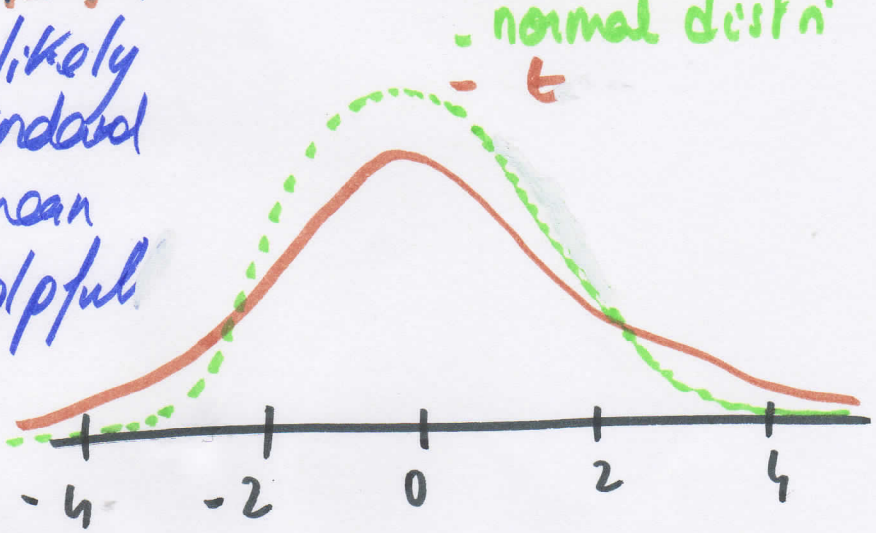
Margin of error: distance required to travel in either direction away from the point estimate when constructing a confidence interval.

Calculating p-value: area under the normal curve beyond the observed sample mean (either in one tail or both). You can use a Z score, where

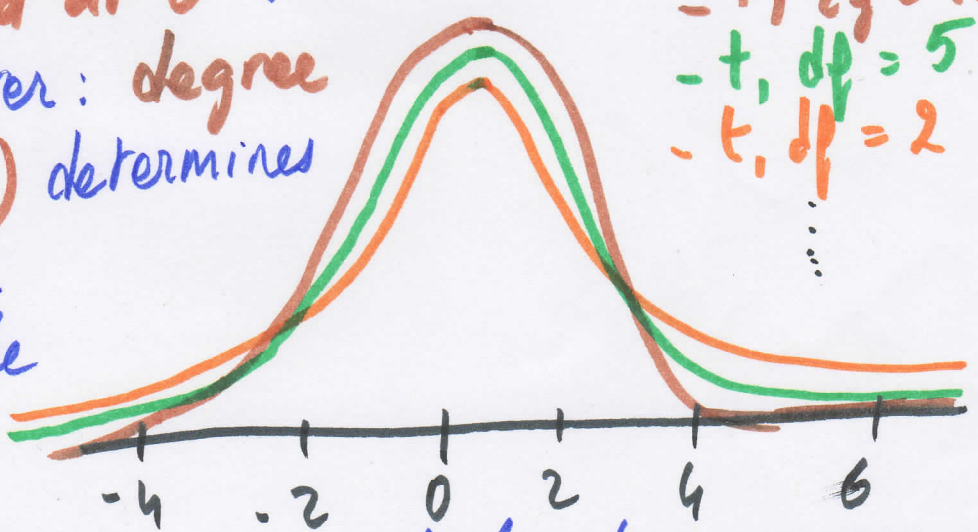
$$Z = \frac{\text{sample statistic} - \text{null value}}{SE} = \frac{\bar{x} - \mu_0}{SE}$$

t-distribution

- When σ **unknown** (almost always), use the **t-distribution** to address the uncertainty of the standard error estimate
- Bell shaped but **fatter tails** than the normal distribution
  - Observations more likely to fall beyond 2 standard deviations from the mean
  - Extra thick tails helpful for mitigating the effect of a less reliable estimate for the standard error of the sampling distribution.



- Always **centered at 0** (like standard normal)
- Has one parameter: **degree of freedom (df)** determines thickness of tails.
- The shape of the t distribution approaches the normal distribution as degrees of freedom increases



t statistic

- For inference on a mean where σ **unknown** (which is most always)

DATA ANALYSIS & STATISTICAL INFERENCE

Duke

(32) Unit 4

→ • calculated the same way $T = \frac{obs - null}{SE}$

• p-value (some definition)
  • one or two tail area, based on H_A
  • using R, applet or table

Examples of calculation

• $P(|t_{df=50}| > 2) = 0.0509$

In R: $pt(2, df = 50, lower.tail = FALSE) \times 2$

• $P(|t_{df=10}| > 2) = 0.0734$

• As we get more conservative with a t-distribution having a low degree of freedom, we become less likely to reject the null hypothesis

t-distribution inference for a mean

Estimating the mean

point estimate ± margin of error

$$\bar{x} \pm t^*_{df} SE_{\bar{x}}$$

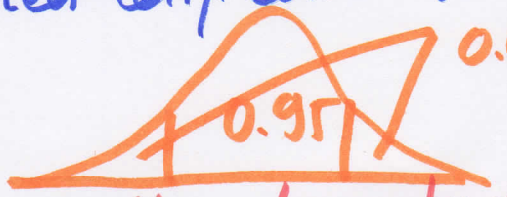
$$\bar{x} \pm t^*_{df} \frac{s}{\sqrt{n}}$$

$$\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

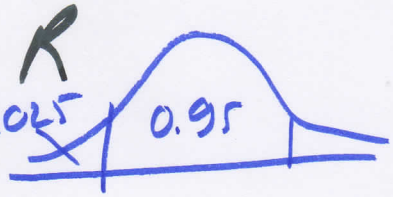
Degrees of freedom for t statistic for inference on one sample mean $df = n - 1$

Finding the critical t-score - using the table

- 1. Determine df and the row
- 2. Find corresponding tail area for desired confidence level



Finding the critical t-score - Using R



- $qt(0.025, df = 21) = -2.07$

Inference for a mean. t-distribution. Example

$\bar{x} = 52.1g$ $s = 45.1g$ $n = 22$ $t^*_{21} = 2.08$

1. Estimate the average after-lunch snack consumption (in grams) of people who eat lunch distracted using a 95% confidence interval.

$$\bar{x} \pm t^* SE = 52.1 \pm 2.08 \times \frac{45.1}{\sqrt{22}}$$

$$= 52.1 \pm 2.08 \times 9.62 \text{ SE}$$

$$= 52.1 \pm 20 = (32.1, 72.1)$$

2. Suppose the suggested serving size of these biscuits is 30g. Do these data provide convincing evidence that the amount of snacks consumed by distracted eaters post-lunch is different than the suggested serving size?

$H_0: \mu = 30$

$H_A: \mu \neq 30$

$$T = \frac{52.1 - 30}{9.62} = 2.3$$

DATA ANALYSIS & STATISTICAL LEARNING

Duke

(34) Unit 4 (part 1 (2) (3))

→ df = 22 - 1 = 21

- Find the p-value (see below)

p-value ≈ 0.0318

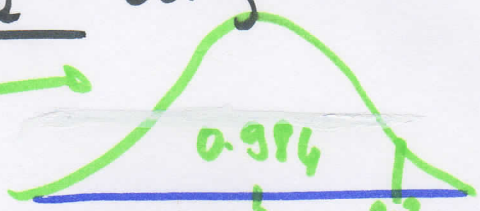
We reject H_0 ($\mu = 30$)

- The two methods (confidence interval & p-value) agree

hypothesis

Finding the p-value - using R

- pt(2.3, df = 21) = 0.984



- (example above) 2 * pt(2.3, df = 21, lower.tail = FALSE)

Finding the p-value - using the table

1. Determine df
2. Locate the calculated T score in the df row
3. Grab the one or two tail p-value from the top row

... < p-value < ...

One tail	0.1	...	0.005
Two tail	0.2	...	0.010
df	1		
	2		
	⋮		
	27		

Inference for comparing two independent means

Estimating the difference between independent means

point estimate ± margin of error

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} SE(\bar{x}_1 - \bar{x}_2)$$

→A

(35) Unit 4 Part 1 (3)

→ Standard error of difference between two independent means $SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Degree of freedom for t statistic for inference on difference of two means $df = \min(n_1 - 1, n_2 - 1)$

Conditions for inference for comparing two independent means

1. Independence
   - within groups: sampled observations must be independent
     - random sample/assignment
     - if sampling without replacement, n < 10% of pop.
   - between groups: the two groups must be independent of each other
2. Sample size/skew: The more skew in the population distributions, the higher the sample size needed.

Examples for inference for comparing two independent means

biscuit intake	$\bar{x}$	s	n
solitaire	52.1g	45.1g	22
no distraction	27.1g	26.4g	22

1. Estimate the difference between the average post-meal snack consumption between those who eat with and without distraction.

$(\bar{X}_{wd} - \bar{X}_{wod}) \pm t_{df}^* SE = (52.1 - 27.1) \pm 2.08 \times \sqrt{\frac{45.1^2}{22} + \frac{26.4^2}{22}}$

(36) Unit 4 Part 1 (3) (4)

→ $= 25 \pm 2.08 \times 11.14 = 25 \pm 23.17$

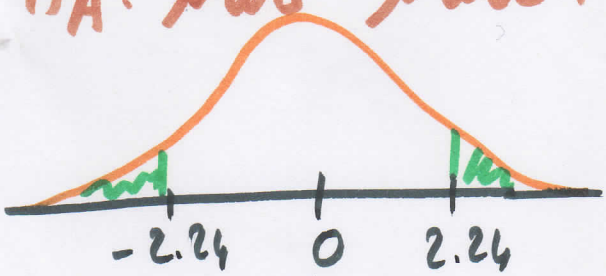
We are 95% confident that those who eat with distractions consume between 1.83g and 48.17g more snacks than those who eat without distractions, on average.

---

2. Do these data provide convincing evidence of a difference between the average post-meal snack consumption between those who eat with and without distractions?

(The hypotheses are always about the population)

$H_0: \mu_{wd} - \mu_{wod} = 0$ $\quad H_A: \mu_{wd} - \mu_{wod} \neq 0$



$T_{21} = \frac{25 - 0}{11.14} = 2.24$

p-value $\simeq 0.036$ ← n-1

pt(2.24, df = 21, lower.tail = FALSE) * 2

We reject H_0. It provides evidence of a difference → The results of the two methods agree

---

Analyzing paired data

- When two sets of observations have this special correspondence (not independent), they are said to be paired
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations.

diff = 1st set − 2nd set

(37) Unit 4 Part 1 (4)

Parameter of interest

Average difference between the 1st set and 2nd set of obs. of **all** population.

μ_{diff}

Point estimate

Average difference between the 1st set and 2nd set of obs. of **sampled** population

$\bar{x}_{diff}$

Inference for comparing two paired means - Example

① One numerical variable

② Hypothesis about the mean

$H_0: \mu_{diff} = 0$

$H_A: \mu_{diff} \neq 0$

③ Calculate the test statistic and the p-value for this hypothesis test

$H_0: \mu_{diff} = 0$

$H_A: \mu_{diff} \neq 0$

$\bar{x}_{diff} = -0.545$

$s_{diff} = 8.887$

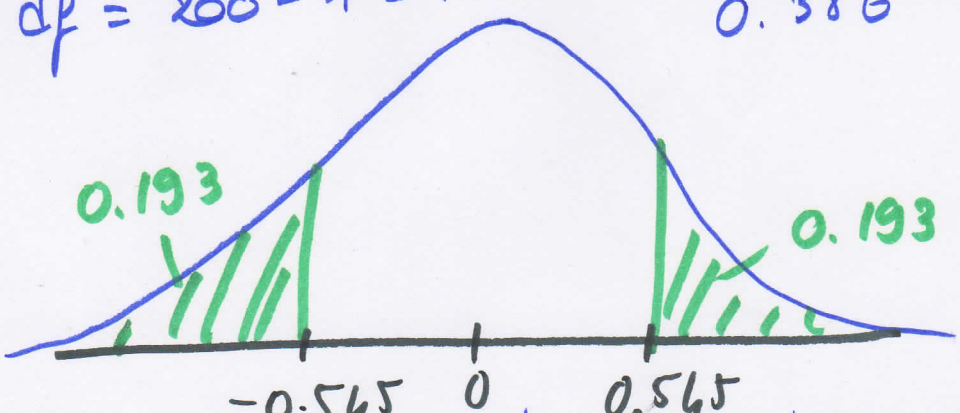
$n_{diff} = 200$

$$T = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} = -0.87$$

$df = 200 - 1 = 199$

p-value = 0.193×2

0.386



- Fail to reject the null hypothesis, the data do not provide convincing evidence of a difference between the average reading and writing scores.

(38) Unit 4 Part 1 (4) Part 2

Summary of inference for comparing two paired means

- Paired data (2 variables) → Difference (1 variable)
- Most often $H_0: \mu_{diff} = 0$
- Same individuals: pre-post studies, repeated measures..
- Different (but dependent) individuals: twins, partners, etc

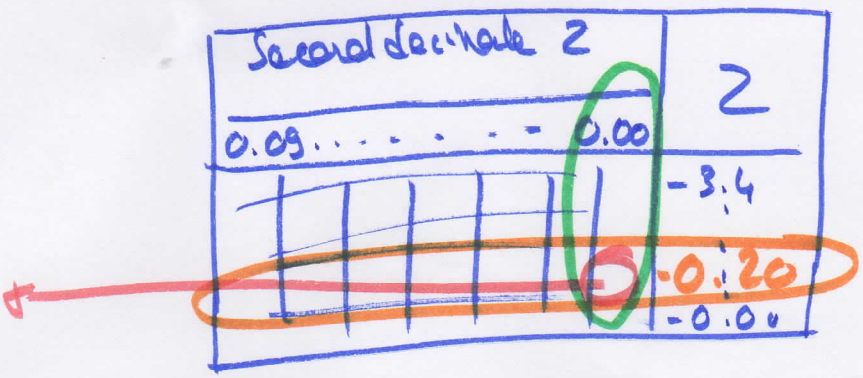
---

Power: power of a test is the probability of correctly rejecting H_0, which is $1 - \beta$ (look (27) Decision errors)

() How to find the area under the normal curve below

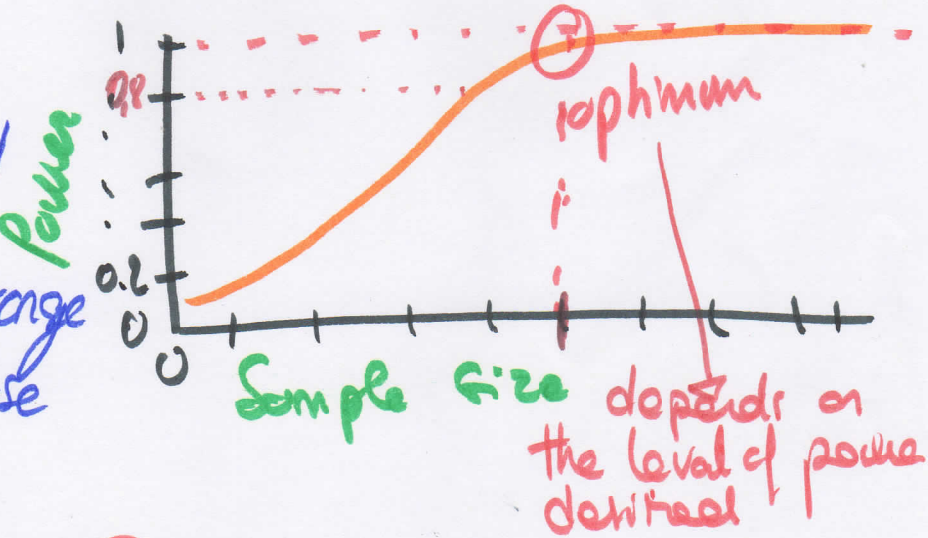
$Z = -0.20$

- Look at the Normal probability table
- area under the curve



Power

- Calculate required sample size for a desired level of power.
- Calculate power for a range of sample sizes, and choose target power.



Following PART 2 (48)

ANOVA (Analysis of variance) Comparing more than two means

H_0: The mean outcome is the same across all categories

μ_i: mean of the outcome for observations in category i

k: number of groups. $\mu_1 = \mu_2 = \dots = \mu_k$

H_A: At least one pair of means are different from each other.

---

t-test

- Comparing means from **two** groups: are so far apart that the observed difference cannot reasonably be attributed to sampling variability

$$H_0: \mu_1 = \mu_2$$

- Compute a test statistic (ratio)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE(\bar{x}_1 - \bar{x}_2)}$$

ANOVA

- Compare means from **more than two** groups: are they so far apart that the observed differences cannot all reasonably be attributed to sampling variably.

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

- Compute a test statistic (ratio)

$$F = \frac{\text{variability between groups}}{\text{variability within groups}}$$

- Large test statistics lead to small p-values
- If the p-value is small enough H_0 is rejected, and we conclude that the data provide evidence of a difference in the population means.

DATA ANALYSIS & STATISTICAL INFERENCE Duke

Unit 4 Part 3 (1) (3)

$$F = \frac{\text{variability between groups}}{\text{variability within groups}}$$

- In order to be able to reject H_0, we need a small p-value, which requires a large F statistic.
- Obtaining a large F statistic requires that the variability between sample means is greater than the variability within the samples.

Conditions for ANOVA (summary)

1. Independence:
   - within groups: sampled observations must be independent.
   - between groups: groups must be independent of each other (non-paired)
2. Approximate normality: distribution should be nearly normal within each group.
3. Equal variance: groups should have roughly equal variability

Conditions for ANOVA (detailed)

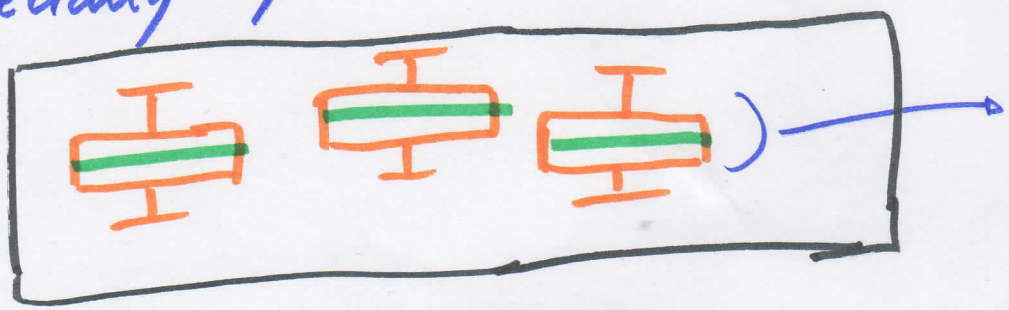
1. Independence

Within:
- Random sample / assignment
- Each n_j less than 10% of respective population
- Always important, but sometimes difficult to check

DATA ANALYSIS & STATISTICAL INFERENCE — Duke

(41) Unit 4 Part 3 (3) (2)

1. **Independence between**: carefully consider whether the groups may be dependent. **Repeated measures ANOVA**
2. **Approximately normal**: distribution of response variable within each group should be approximately normal
   - Especially important when sample sizes are small.
3. **Constant variance**: Variability should be consistent across groups **homoscedastic** groups
   - Especially important when sample size differ between groups



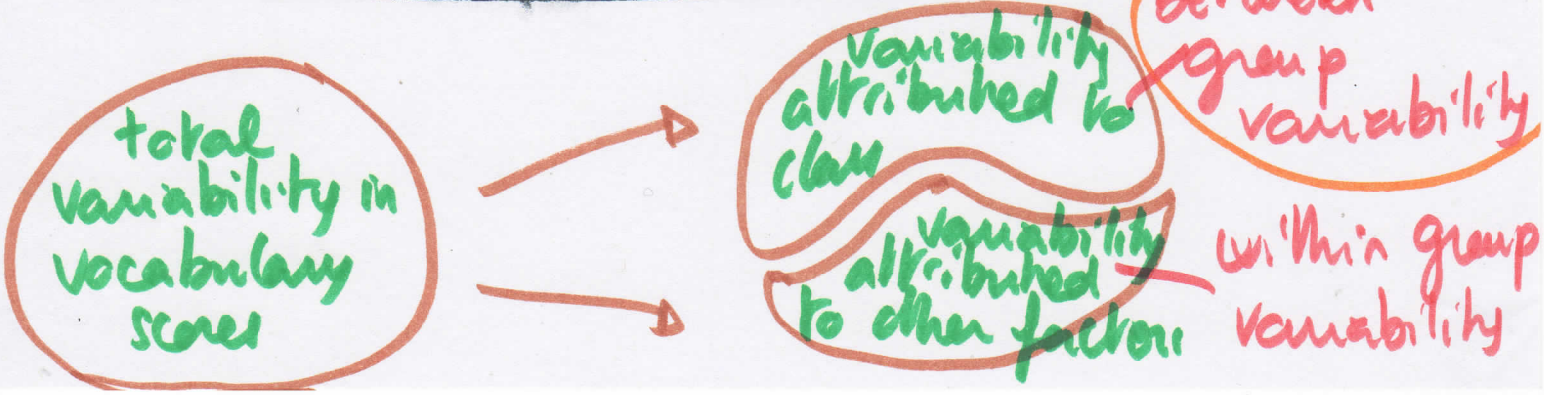
- In ANOVA, the numerical variable is the response variable.

ANOVA - Variability partitioning (with example)

Response variable: Vocabulary score
Explanatory variable: social class

	Vocabulary score	social class
1		
...		
795		

We are interested in this variability



DATA ANALYSIS & STATISTICAL INFERENCE Duke

(42) Unit 4 Part 3(2)

ANOVA Output

	Degree of freedom	Sum of Squares	Mean Squares	F value	Pr(>F)
Group	df_G	SSG	MSG	F	p-value
Error	df_E	SS_E	MS_E		
Total	df_T	SST	MST		

Group: between group variability
Error: within group variability

What is interesting is how it is splitted between SSG & SSE

SST = measures the **total variability** in the response variable (Sum of squares Total)

- Calculated very similarly to variance [except not scaled by the sample size]

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

y_i = value of the response variable for each observation

$\bar{y}$ = grand mean of the response variable

SSG (Sum of squares groups): measures the variability between groups.

- Variability of the response variable explained by the explanatory variable

$$SSG = \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2$$

n_j = nb of obs. in group j
$\bar{y}_j$ = mean of the response var. for group j
$\bar{y}$ = grand mean of response var.

DATA ANALYSIS & STATISTICAL INFERENCE

Dike

(43) Unit 4 Part 3 (2)

SSG: On its own, it's not a meaningful number. It is interesting to see how it compared to SST $X = \frac{SSG}{SST}$

X % of the variability in the response variable is explained by the explanatory variable

SSE (Sum of squares error): measures the variability within groups. Unexplained variability: unexplained by the group variable, due to other reasons.

$SSE = SST - SSG$

Degrees of freedom associated with ANOVA

$df_T = n - 1$

$df_G = K - 1$

$df_E = df_T - df_G$

Mean Squares: Average variability between and within groups, calculated as the total variability (sum of squares) scaled by the associated degrees of freedom.

$MSG = SSG / df_G$

$MSE = SSE / df_E$

F statistic: Ratio of the average between group and within group variabilities.

$F = \frac{MSG}{MSE}$

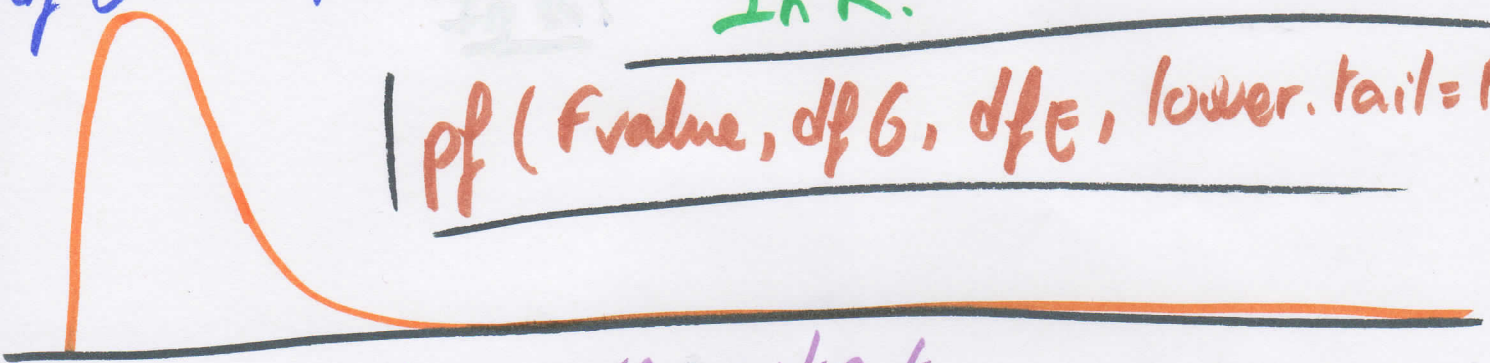
DATA ANALYSIS & STATISTICAL INFERENCE

Duke

Unit 4 Part 3(2)

(same but explained differently)

p-value: • probability of at least as large a ratio between the "between" and "within" group variabilities if in fact the means of all groups are equal.
• Area under the F curve, with degrees of freedom df_G and df_E, above the observed F statistic.



In R:

```
pf(Fvalue, dfG, dfE, lower.tail=F)
```

Conclusion of ANOVA output

- If p-value is small (less than α), reject H_0
  - The data provide convincing evidence that at least one pair of population means are different from each other (but we can't tell which one)
- If p-value is large, fail to reject H_0
  - The data do not provide convincing evidence that at least one pair of population means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance)

Multiple comparisons

- Testing many pairs of groups

DATA ANALYSIS & STATISTICAL INFERENCE

Duke

(45) Unit 4 Part 3(4)

→ • The **Bonferroni correction** suggests that a more stringent (≃ rigorous) significance level is more appropriate for these tests

• adjust α by the number of comparisons being considered?

Bonferroni correction

K = number of comparisons

$$\alpha^* = \alpha / K \qquad K = \frac{k(k-1)}{2}$$

Bonferroni correction - Example

The social class variable has **4 levels**. If $\alpha = 0.05$ for the original ANOVA, what should the modified significance level be for two sample t tests for determining which pairs of groups have significantly ≠ means?

$k = 4$ $\quad K = \frac{4 \times 3}{2} = 6$ comparisons

$\alpha^* = 0.05 / 6 \simeq 0.0083$ (new significance level for the pairwise comparisons)

Pairwise comparisons

- Constant variance → Re-think standard error and degrees of freedom: use consistent standard error and degrees of freedom for all tests
- Compare the p-values from each test to the modified significance level.

DATA ANALYSIS & STATISTICAL INFERENCE

Duke

(46) Unit 4 Part 3(4) Part 4(1)

Standard error for multiple pairwise comparison

$$SE = \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

(multiple pairwise comparisons)

Degrees of freedom for multiple pairwise comparison

$$df = df_E$$

Statistical Inference: we use a sample of a population to make inference on that population

---

Bootstrapping scheme

(1) Take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample

(2) Calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples.

(3) Repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics

---

Bootstrap vs. sampling distribution

- Sampling distribution created using sampling (with replacement) from the population

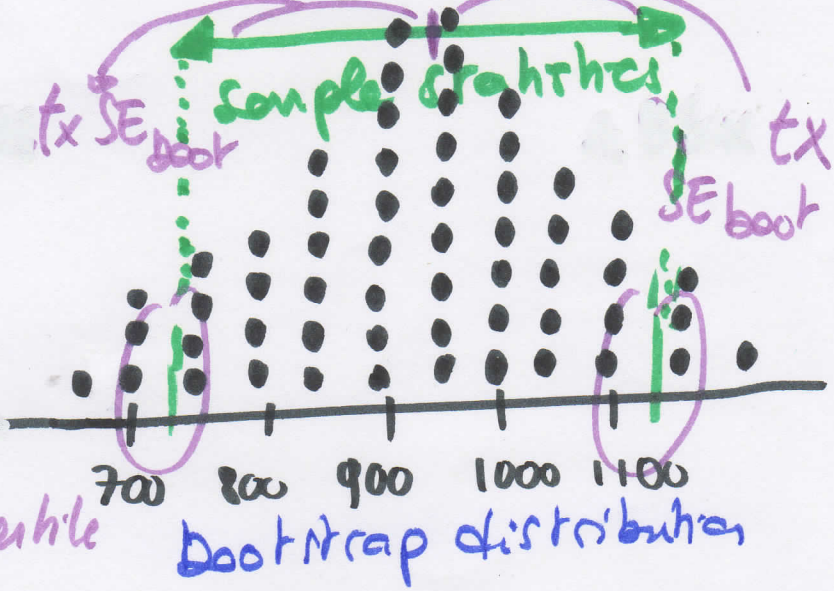
(47) Unit 4 Part 4 (1)

- Bootstrap distribution created using sampling (with replacement) from the sample
- Both are distributions of sample statistics

Bootstrapping - Calculating Confidence Interval 2 ways

① Percentile method

② Standard error method (more accurate)



If the statistic calculated is the median, then that point is a median from a bootstrap sample

Bootstrapping limitations

- Not as rigid conditions as CLT based methods
- If the bootstrap distribution is extremely skewed or sparse, the bootstrap interval might be unreliable
- A representative sample is still required - if the sample is biased, the estimates resulting from this sample will also be biased.

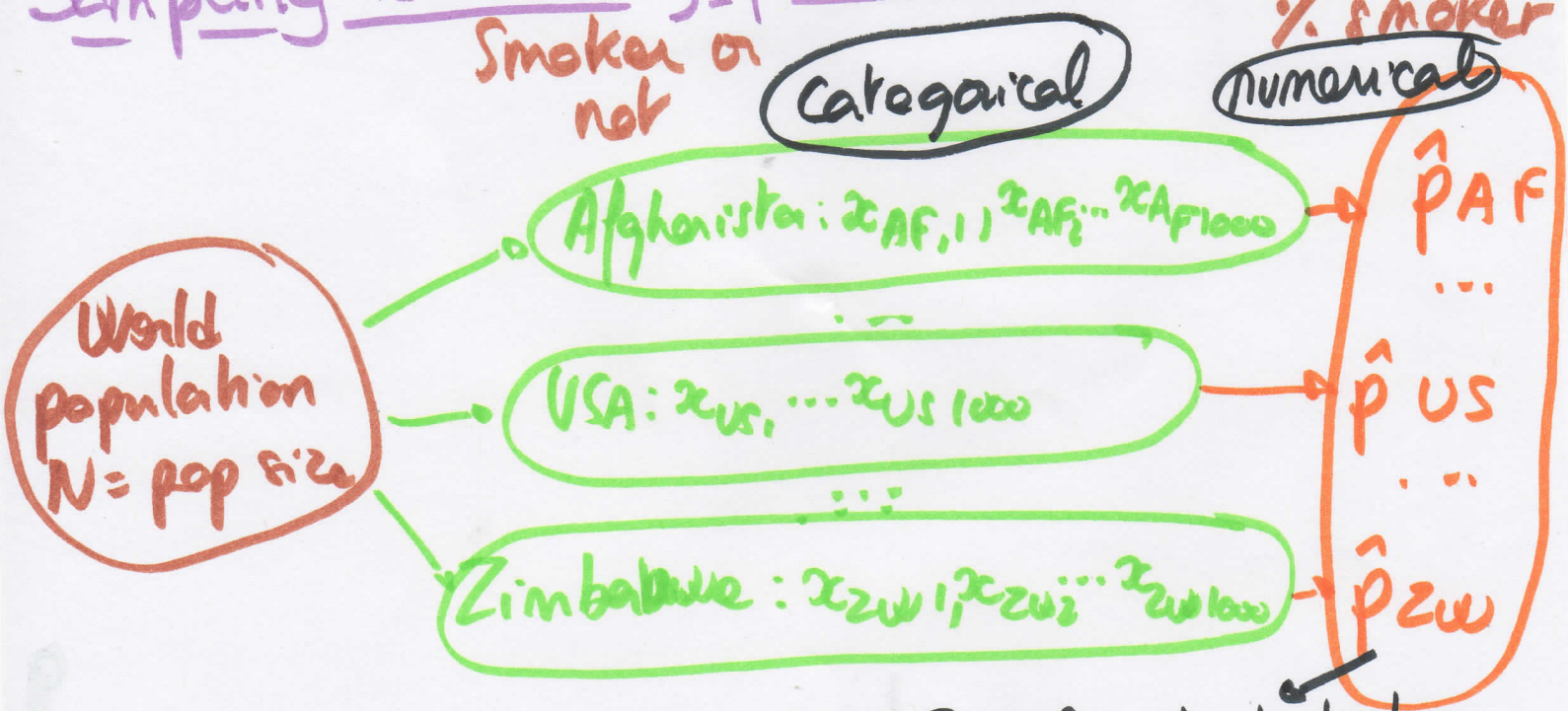
DATA ANALYSIS & STATISTICAL ANALYSIS Done

(48) Unit 4 Part 2 following (38) + Unit 5 Part 1 (1)

Power

In an hypothesis test, our goal is to keep α and β low. (27)
However, decreasing one increases the other (see decision error)
One solution to this problem is increasing our sample size.
We have to obtain a sufficiently large sample size to have the required power.
(+ Example see video)

Sampling variability for proportions - Example



$$p = \frac{\#\text{ of smokers in the world}}{N}$$

$$\text{mean}(\hat{p}) \simeq p$$

CLT for proportions: The distribution of sample proportions is nearly normal, centered at the population proportion, and with a standard error inversely proportional to the sample size

DATA ANALYSIS & STATISTICAL INFERENCE Duke

(49) Units Part 1 (1)

CLT for proportions

$$\hat{p} \sim N\left(\text{mean} = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

shape — center — spread

Conditions for the CLT for proportions

1. **Independence**: Sampled observations must be independent
   - Random sample/assignment
   - If sampling without replacement, $n < 10\%$ of pop.
2. **Sample size/skew**: There should be at least 10 successes and 10 failures in the sample: $np \geq 10$, $n(1-p) \geq 10$

If p unknown, use $\hat{p}$ (sample proportion)

CLT for proportions - Example

90% of all plants species are classified as angiosperms (flowering plants). If you were to randomly sample 200 plants from the list of all known plant species, what is the probability that at least 95% of plants in your sample will be flowering plants. $p = 0.90$ $n = 200$

$P(\hat{p} > 0.95) = ?$

1. random sample & $< 10\%$ of all plants → independent obs.
2. $200 \times 0.90 = 180$ and $200 \times 0.10 = 20$

CLT for proportion example (following)

$\hat{p} \sim N(\text{mean} = 0.90, SE = \sqrt{\frac{0.90 \times 0.10}{200}} \simeq 0.0212)$

$Z = \frac{0.95 - 0.90}{0.0212} = 2.36$

$P(Z > 2.36) \simeq 0.0091$ — very close — 2 different methods

Using the binomial distribution: $200 \times 0.95 = 190$

In R: sum(dbinom(190:200, 200, 0.90)) # 0.00802

What if the success-failure condition is not met?

- The center of the sampling distribution will still be around the true population proportion.
- The spread of the sampling distribution can still be approximated using the same formula for the standard error.
- The shape of the distribution will depend on whether the true population proportion is closer to 0 or closer to 1 — right skew / left skewed

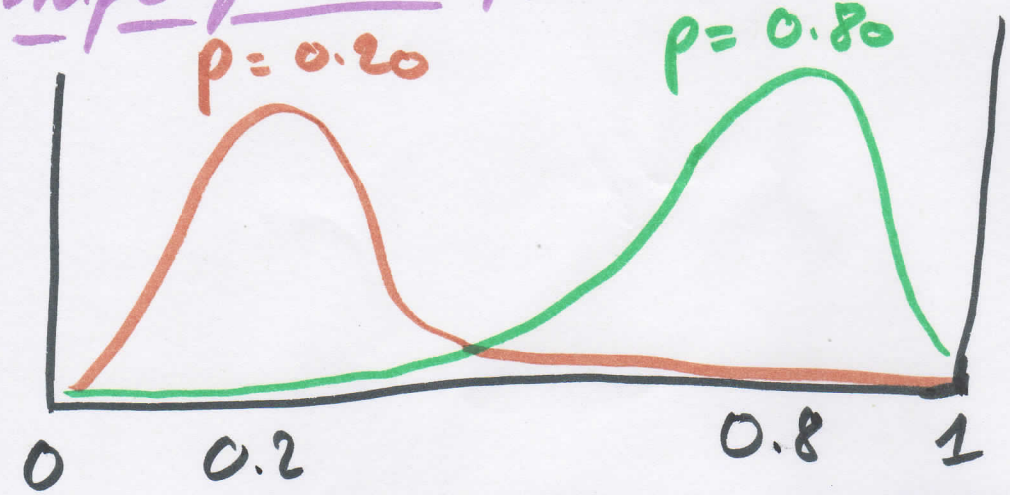
---

A proportion considered usual: within two standard errors. Use Z to calculate it.

DATA ANALYSIS & STATISTICAL INFERENCE

(51) Unit 5 Part 1 (1) (2) Duke

Shape of the sampling distribution



parameter of interest: p

Estimating a proportion

point estimate ± margin of error

$$\hat{p} \pm z^* SE\,\hat{p}$$

Standard error for a proportion for calculating a confidence interval:

$$SE\,\hat{p} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence interval for proportion - example

85% of Americans have good intuition about experimental design (GSS). Estimating (using 95% confidence interval) the proportion of all Americans who have this good intuition

- CLT conditions met (independence / sample size - skew)

$$\hat{p} \pm z^* SE = 0.85 \pm 1.96\sqrt{\frac{0.85 \times 0.15}{670}}$$

$$= (0.823, 0.877)$$

(52) Unit 5 Part 1 (2) (3)

→ We are 95% confident that 82.3% to 87.7% of all Americans have good intuition about experimental design.

Calculating the required sample size for desired ME

Remember $ME = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$$n = \frac{z^{*2} \times \hat{p} \times (1-\hat{p})}{ME^2 \text{ (desired ME)}}$$

- If there is a previous study that we can rely on for the value of $\hat{p}$, use that in the calculation of the required sample size.
- If not, use $\hat{p} = 0.5$
  - if you don't know any better, 50-50 is a good guess
  - gives the most conservative estimate - highest possible sample size

Hypothesis testing for a single proportion

1. Set the hypotheses: H_0: p = null value
   H_A: $p <$ or $>$ or $\neq$ null value
2. Calculate the point estimate: $\hat{p}$
3. Check conditions: Independence & sample size/skew
4. Draw sampling distribution, shade p-value, calculate test statistic $Z = \frac{\hat{p} - p}{SE}$, $SE = \sqrt{\frac{p(1-p)}{n}}$

Unit 5 Part 2 (3)

5- Make a decision, and interpret it in context of the research question:

- If p-value $< \alpha$, reject H_0; the data provide convincing evidence for H_A
- If p-value $> \alpha$ fail to reject H_0 the data do not provide convincing evidence for H_A

$\hat{p}$ vs. p

One proportion

	Observed Confidence interval	Expected hypothesis test
Success-failure condition	$n\hat{p} \geq 10$ $n(1-\hat{p}) \geq 10$	$np \geq 10$ $n(1-p) \geq 10$
Standard error	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{p(1-p)}{n}}$

Hypothesis testing for a proportion - Example

A 2013 Pew Research poll found that 60% of 1,983 randomly sampled American adults believe in evolution. Does this provide convincing evidence that majority of Americans believe in evolution?

$H_0: p = 0.5$ $H_A: p > 0.5$ $\hat{p} = 0.6$ $n = 1983$

Independence + sample skew/size met

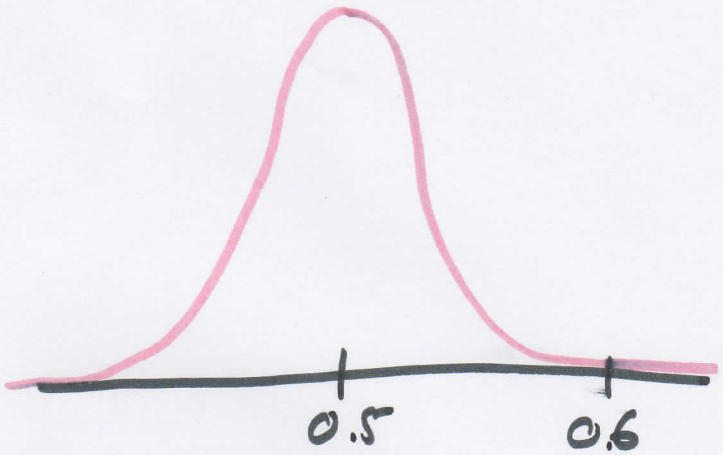
1983 < 10% of Americas. $1983 \times 0.5 = 991.5 > 10$

DATA ANALYSIS & STATISTICAL INFERENCE Duke

(5) Units Part 1 (3)

→ $\hat{p} \sim N(\text{mean} = 0.5, SE = \sqrt{\frac{0.5 \times 0.5}{1983}} \simeq 0.0112)$

$$Z = \frac{0.6 - 0.5}{0.0112} \simeq 8.92$$



p-value = $P(Z > 8.92)$
= almost 0

Reject H_0

- There is almost 0% chance of obtaining a random sample of 1,983 Americans where 60% or more believe in evolution, if in fact 50% of americans believe in evolution.

Estimating the difference between two proportion

point estimate ± margin of error

$$(\hat{p}_1 - \hat{p}_2) \pm z^* SE_{(\hat{p}_1 - \hat{p}_2)}$$

Standard error for difference between two proportions, for calculating a confidence interval:

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Conditions for inference for comparing two independent proportions

1. Independence:

- Within groups: sampled observations must be independent within each group
  - random sample / assignment
  - if sampling without replacement, n < 10% of population
- Between groups: the two groups must be independent of each other (non-paired)

2. Sample size / skew: Each sample should meet the success-failure condition:
   - $n_1 p_1 \geq 10$ and $n_1(1-p_1) \geq 10$
   - $n_2 p_2 \geq 10$ and $n_2(1-p_2) \geq 10$

Example

Estimating the difference between two proportions

Parameter of interest: $p_{Coursera} - p_{US}$

Point estimate: $\hat{p}_{Coursera} - \hat{p}_{US}$

Using a 95% confidence interval, estimate how Coursera students and the American public at large compare with respect to their views on laws banning possession of handguns.

	Success	n	$\hat{p}$
US	257	1028	0.25
Coursera	59	83	0.71

1. independence: random sample: yes for US, no for Coursera
10% condition: met for both.

⚠ Sampled Courserians may not be independent of each other

DATA ANALYSIS & STATISTICAL INFERENCE Duke

(57) Unit 5 Part 2 (2)

$\hat{p}$ vs. $\hat{p}$ Revisit: working with two proportion

	Observed Confidence interval	Expected Hypothesis test
success-failure condition	$n_1\hat{p}_1 \geq 10$ $n_1(1-\hat{p}_1) \geq 10$ $n_2\hat{p}_2 \geq 10$ $n_2(1-\hat{p}_2) \geq 10$	$n_1\hat{p}_{pool} \geq 10$ $n_1(1-\hat{p}_{pool}) \geq 10$ $n_2\hat{p}_{pool} \geq 10$ $n_2(1-\hat{p}_{pool}) \geq 10$
standard error	$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$SE = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$

Small Sample proportion (condition sample size not met)

Inference via simulation (revisit).

- The ultimate goal of a hypothesis test is a p-value
  - p-value = P(observed or more extreme outcome | H_0 true)
- Devise a simulation scheme that assumes the null hypothesis is true
- Repeat the simulation many times and record relevant sample statistic
- Calculate p-value as the proportion of simulations that

DATA ANALYSIS & STATISTICAL INFERENCE Duke

(56) Unit 5 Part 2 (1) (2)

→ Sample size / skew US: 257 successes, 1028 − 257 = 771 failures

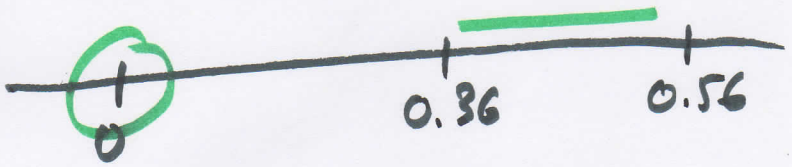
Coursera: 59 successes, 83 − 59 = 24 failures

$$(\hat{p}_{coursera} - \hat{p}_{US}) \pm z^* SE = (0.71 - 0.25) \pm 1.96 \sqrt{\frac{0.71 \times 0.29}{83} + \frac{0.25 \times 0.75}{1028}}$$

$$(p_{coursera} - p_{US}) = (0.36, 0.56)$$

Based on the CI we calculated, should we expect to find a significant difference (at the equivalent significance level) between the population proportion of Coursera students and the American public at large who believe there should be a law banning the possession of handguns?

$H_0: p_{coursera} - p_{US} = 0$



We reject H_0

Pooled proportion

$H_0: p_1 = p_2 = ?$

$\hat{p}_{pool} = \frac{\text{total successes}}{\text{total } n}$

Pooled proportion: $= \frac{\#\text{ of successes}_1 + \#\text{ of successes}_2}{n_1 + n_2}$

L6 for comparing two proportions

(58) Unit 5 Part 3 (1) (2) (3) Part 4 (1)

→ yield a result favorable to the alternative hypothesis

In R: inference function method = "simulation"

Small Sample proportion - Example

About the saying in English "to know something like the back of your hand"

In R: source("http://bit.ly/dasi_inference")

back = factor(c(rep("correct", 11), rep("incorrect", 1)))

inference(back, est = "proportion", type = "ht", method = "simulation", success = "correct", null = 0.1, alternative = "greater", nsim = 100)

Chi-Square Goodness of Fit Test

Evaluating the hypotheses

- Quantify how different the observed counts are from the expected counts
- Large deviations from what would be expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis
- Called a **goodness of fit** test since we're evaluating how well the observed data **fit** the expected distribution

Conditions for the Chi-square test

1. **Independence**: Sampled observation must be independent
   - Random sample / assignment

(59) Unit 5 Part 4 (1)

- If sampling without replacement, $n < 10\%$ of population
- Each case only contributes to one cell in the table

2. **Sample size**: Each particular scenario (i.e. cell) must have at least 5 expected cases.

Chi-square statistic

When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the **chi-square (χ^2) statistic**

χ^2 statistic: $$\chi = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

O: observed
E: expected
k: number of cells

Chi-square degrees of freedom

- To determine if the calculated χ^2 statistic is considered unusually high or not we need to first describe its distribution.
- Chi-square distribution has just one parameter:
  - Degrees of freedom (df): influences the shape, center and spread

χ^2 degrees of freedom for a goodness of fit test:

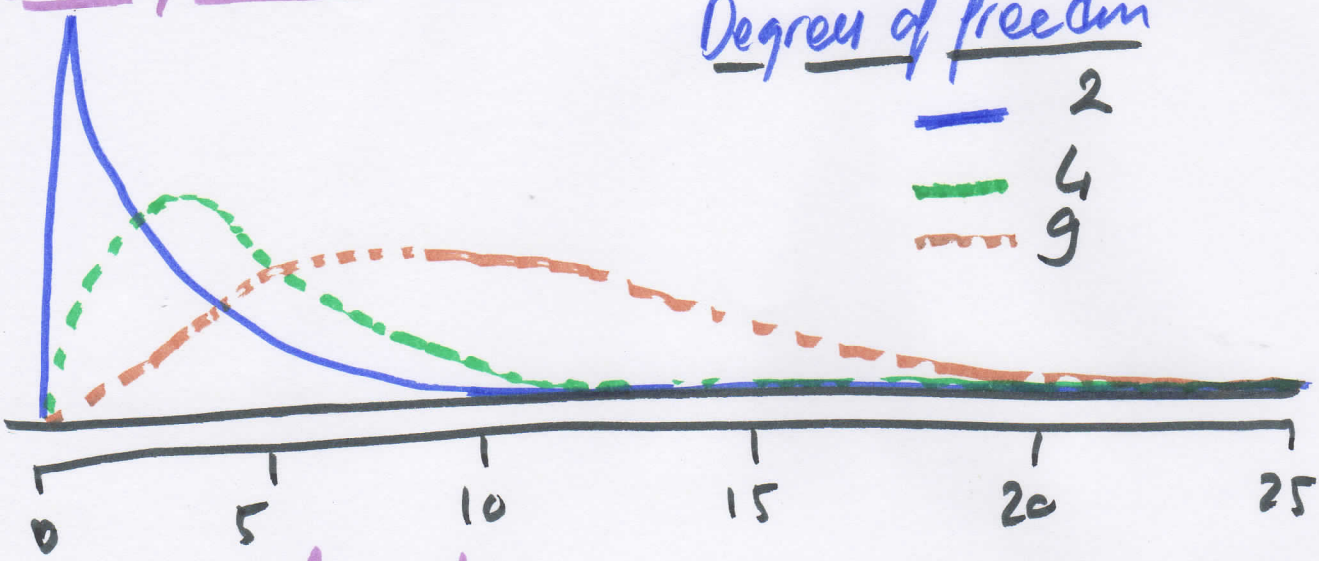
$df = k - 1$

k = number of cells

Chi-square GOF test: one categorical variable, > 2 levels

(60) Unit 5 Part 4 (1)

Chi-square distribution & degrees of freedom



P-value for Chi-square

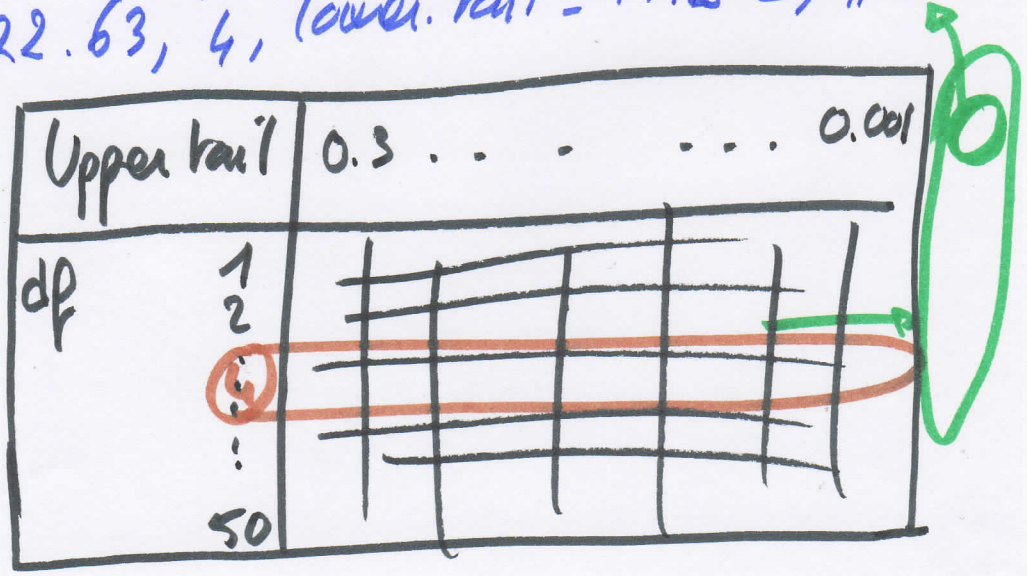
- p-value for a chi-square test is defined as the tail area **above** the calculated test statistic
- Because the test statistic is always positive, and a higher test statistic means a higher deviation from the null hypothesis

Finding p-value for chi-square

Example: $\chi^2 = 22.63$ $df = 4$

Using R: pchisq(22.63, 4, lower.tail = FALSE) = 0.0002

Using the table



DATA ANALYSIS & STATISTICAL INFERENCE

(61) Unit 5 Part 4 (2) Unit 6 Part 1 (1) Duke

Chi-square tests

Goodness of fit: comparing the distribution of one categorical variable (with more than 2 levels) to a hypothesized distribution.

Independence: Evaluating the relationship between two categorical variables (at least one with more than 2 levels).

Chi-square test of independence

X^2 test of independence

$$X^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

O: observed
E: expected
k: number of cells

$$df: (R-1) \times (C-1)$$

R: number of rows C: number of columns

Some conditions (Chi square GOF)

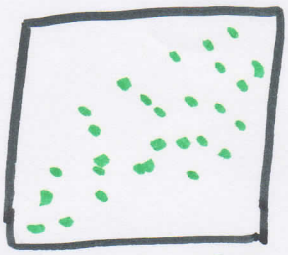
Linear regression-correlation

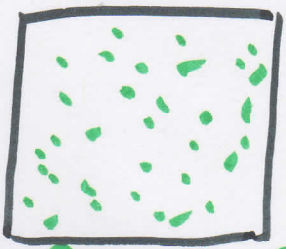
Correlation: describes the strength of the **linear association** between two variables. Denoted as **R**

Correlation properties:

(1) The magnitude (absolute value) of the correlation coefficient measures the strength of the linear association between two numerical variables!


R≃0.97


R≃0.69


R≃0.07

(2) The sign of the correlation coefficient indicates the direction of association

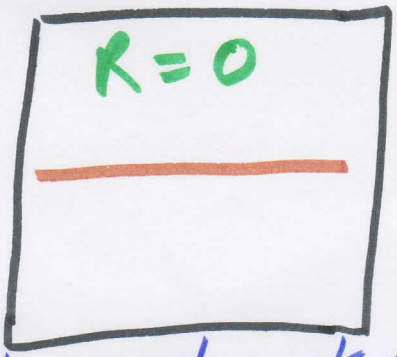

R≃0.98


R≃-0.9

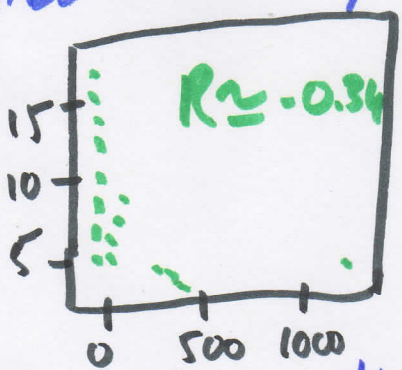
(3) The correlation coefficient is always between -1 (perfect negative linear association) and 1 (perfect positive linear association). R=0 indicates no linear relationship.

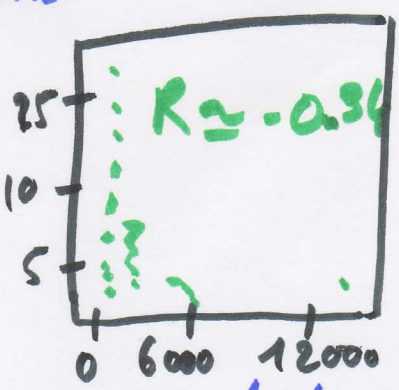

R=1


R=-1

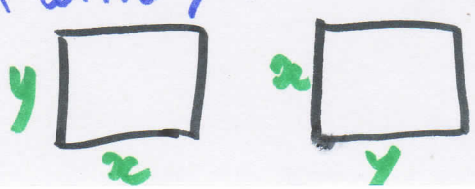

R=0

(4) The correlation coefficient is unitless, and is not affected by changes in the center or scale of either variable (such as unit conversions)





(5) The correlation of X with Y is the same as of Y with X

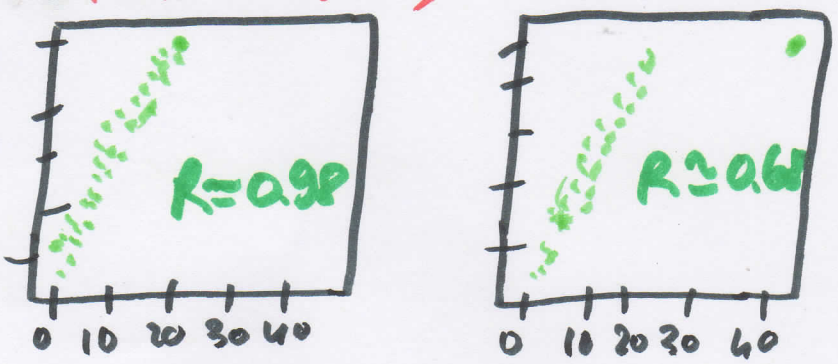


→ same R

DATA ANALYSIS & STATISTICAL INFERENCE Duke

(63) Unit 6 Part 1 (1) Part 2 (1) (2)

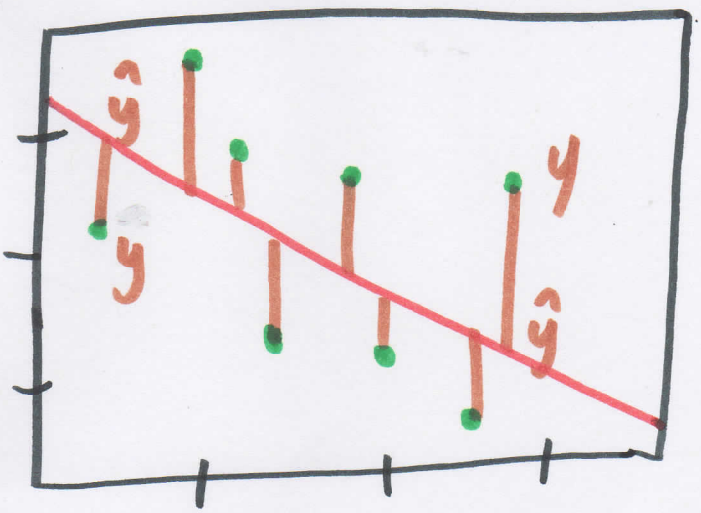
→ (6) The correlation coefficient is sensitive to outliers.



Residuals. Linear regression

- leftovers from the model fit
- data = fit + residual
- difference between the observed and predicted y

Residual: $e_i = y_i - \hat{y}_i$



Least squares line - Linear regression

$\hat{y} = \beta_0 + \beta_1 x$

- $\hat{y}$: predicted response
- β_0: intercept
- β_1: slope
- x: explanatory

notation

	parameter	point estimate
intercept	β_0	b_0
slope	β_1	b_1

DATA ANALYSIS & STATISTICAL INFERENCE Duke

(64) Unit 6 Part 2 (2)

Estimating the regression parameters: slope

s_x: SD of x

s_y: SD of y

$R = cor(x, y)$

slope: $b_1 = \frac{s_y}{s_x} R$

- For each % point increase in explanatory variable rate, we would expect the % response variable to be lower on average by b_1 % points

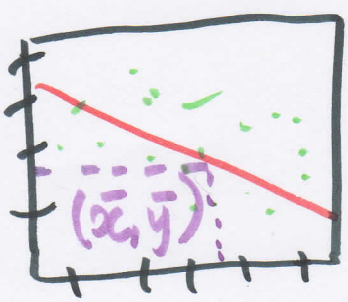
Estimating the regression parameters: intercept

$\hat{y} = b_0 + b_1 x$ ~~x~~ $\bar{x}$

$\bar{y}$

intercept: $b_0 = \bar{y} - b_1 \bar{x}$

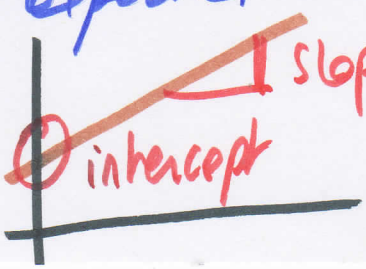
The least squares line always goes through $(\bar{x}, \bar{y})$



Recap intercept & slope - linear regression

Intercept: When $x = 0$, y is expected to equal the intercept. May be meaningless in context of the data, and only serve to adjust the height of the line.

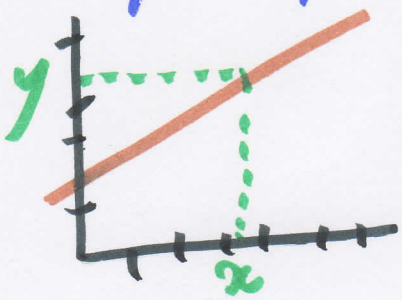
Slope: For each unit increase in x, y is expected to be higher / lower on average by the slope

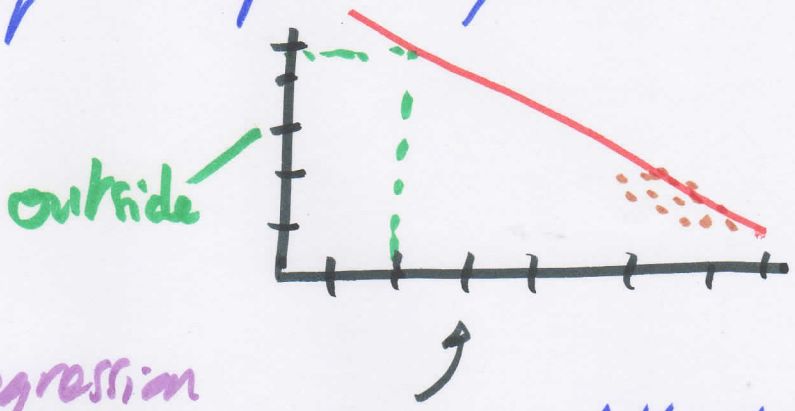


(65) Unit 6 Part 2(3)(4)

Prediction - linear regression

- Using the linear model to predict the value of the response variable for a given value of the explanatory variable





Extrapolation - linear regression

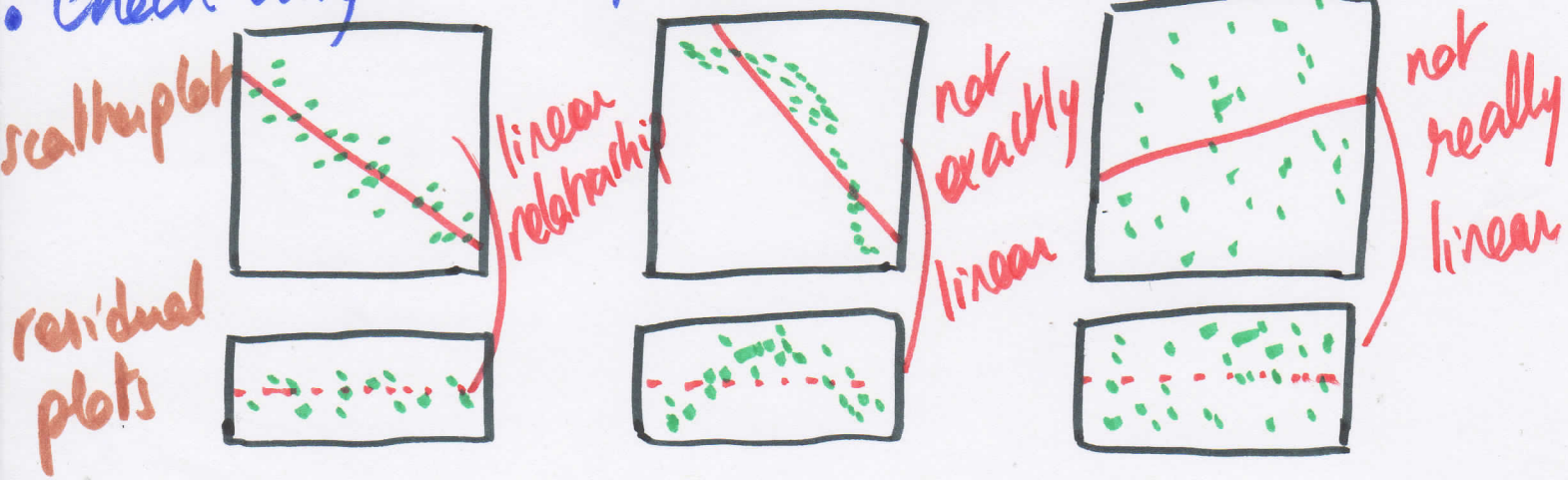
⚠ not always reliable if it is beyond the scope of the data

- Applying a model estimate to values outside of the realm of the original data.

Conditions for linear regression

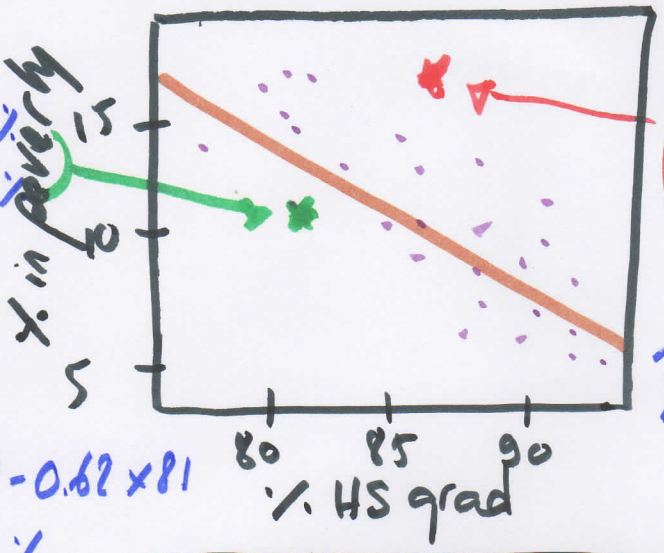
(1) Linearity

- Relationship between the explanatory and the response variable should be linear
- Methods for fitting a model to non-linear relationships exist
- Check using a scatterplot of the data, or a **residual plot**



(66) Unit 6 Part 2 (4)

Anatomy of a residual plot



* R1
% HS grad = 81%
% in poverty = 10.3%

* DC
% HS grad = 86%
% in poverty = 16.8%

$\widehat{\%\ pov} = 64.68 - 0.62 \times 81$
$= 14.46\ \%$

$e = 10.3 - 14.46$
$= -4.16\ \%$

$\widehat{\%\ pov} = 64 - 68 - 0.62 \times 86$
$= 11.36\ \%$

$e = 16.8 - 11.36$
$= 5.44\ \%$

residual plot

x axis = explanatory var
% HS grad

at 0

(2) Nearly normal residuals

- Residuals should be nearly normally distributed, centered
- May not be satisfied if there are unusual observations that don't follow the trend of the rest of the data
- Check using a histogram or normal probability plot of residuals.

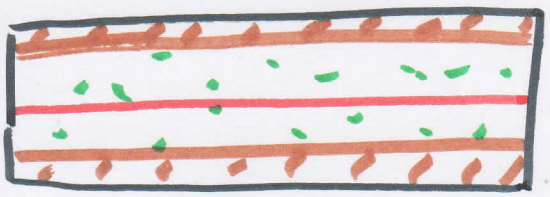
(3) Constant variability

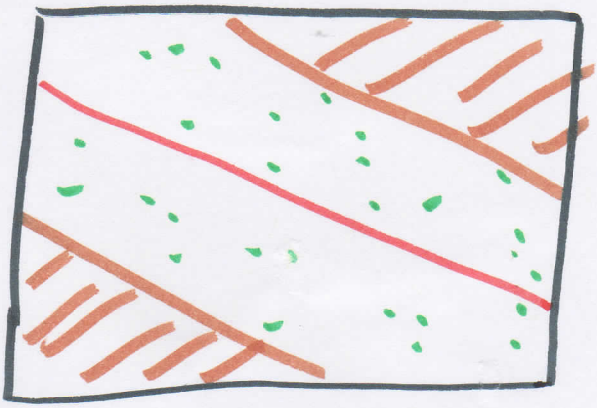
- Variability of points around the least squares line should be roughly constant
- Implies that the variability of residuals around the 0 line should be roughly constant as well
- Also called **homoscedasticity**. Check using a residuals plot

DATA ANALYSIS & STATISTICAL INFERENCE Duke

(67) Unit 6 Part 2(4)(5)(6) Part 3(1)

(Constant variability)





R^2 - linear regression

- Strength of the fit of a linear model is most commonly evaluated using R^2.
- Calculated as the square of the correlation coefficient
- Tells us what percent of variability in the response variable is explained by the model
- The remainder of the variability is explained by variable not included in the model.
- Always between 0 and 1.

Interpretation of the $R^2 (= 0.5625)$ for the model predicting % living in poverty from % HS graduation rate. 56.25% of the variability in the % of residents living in poverty among the states is explained by the model.

Regression with categorical explanatory variable

- Use levels 0 and 1 for the categorical variable
0 for the reference level

Types of outliers

- Outliers are points that fall away from the cloud of points

DATA ANALYSIS & STATISTICAL INFERENCE

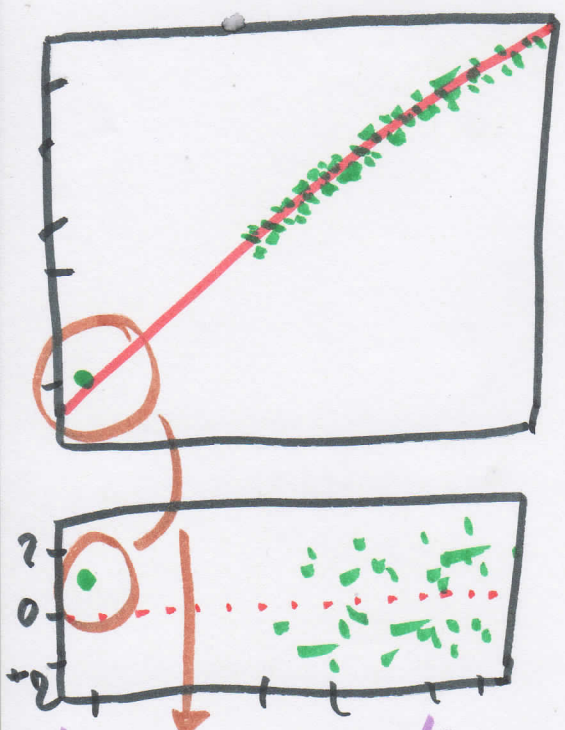
Duke

(68) Unit 6 Part 3(1) Part 4(1)

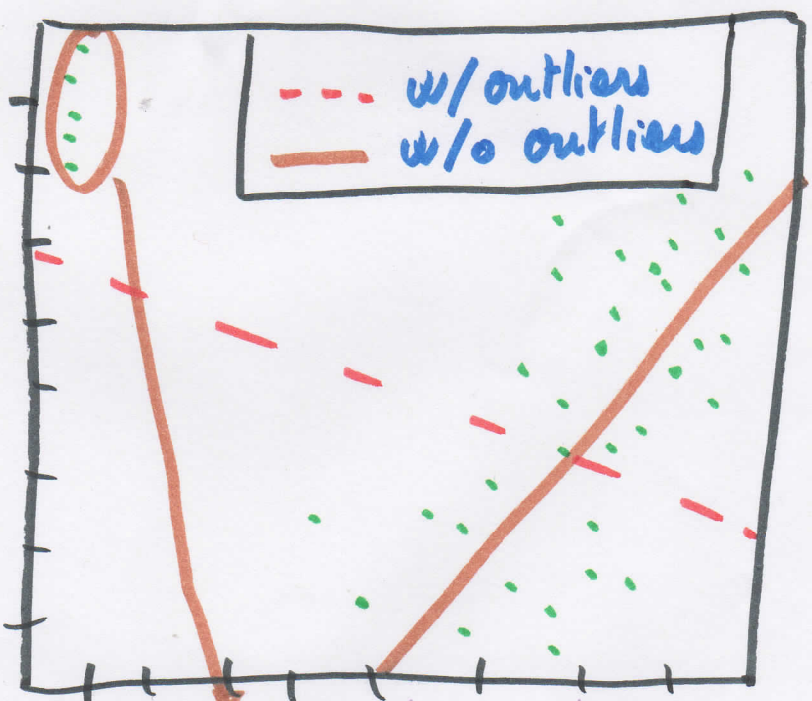
① Leverage points: outliers that fall horizontally away from the center of the cloud but don't influence the slope of the regression line.

② Influential points: outliers that actually influence the slope of the regression line.

- Usually high leverage points
- To determine if a point is influential, visualize the regression line with and without the points, and ask: Does the slope of the line change considerably?



leverage point



Influential points

- Influential points don't always reduce R^2

Inference for linear regression

Testing the slope - hypothesis

$H_0: \beta_1 = 0$. The explanatory variable is not a signi-

69 Unit 6 Part 4 (1)

- Cant predictor of the response variable, i.e. no relationship → slope of the relationship is 0.

H_A $\beta_1 \neq 0$. The explanatory variable is a significant predictor of the response variable, i.e. relationship → slope of the relationship is different than 0.

Use a t-statistic for the slope (hypothesis test)

$$T = \frac{b_1 - 0}{SE_{b_1}} \qquad df = n - 2$$

[df: lose 1 df for each parameter estimated. Here β_0 and β_1]

Confidence interval for the slope

point estimate ± margin of error

$$b_1 \pm t^*_{df} \times SE_{b_1}$$

⚠ However, inference on the intercept is rarely done, mostly on the slope.

Inference for linear regression (slope) - example

Calculate the 95% confidence interval for the slope of the relationship between biological and foster twins' IQs

R output:

	Estimate	Std. Error	t-value	Pr(>|t|)
(intercept)	9.2076	9.2999	0.99	0.3316
bio IQ	0.9014	0.0963	9.36	0.0000

DATA ANALYSIS & STATISTICAL INFERENCE Duke

(20) Unit 6 Part 4 (1) (2)

$df = 27 - 2 = 25$. $t^* = 2.06$

$0.9014 \pm 2.06 \times 0.0963$

$= (0.7, 1.1)$

R:

qt(0.025, df = 25)

-2.059539

Interpretation: We are 95% confident that for each additional point on the biological twins' IQs, the foster twins' IQs (response) are expected on average to be higher by 0.7 to 1.1 points.

Caution!

- Always be aware of the type of data you're working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you already have population data.
- If you have a sample that is non-random (biased), the results will be unreliable.
- The ultimate goal is to have independent observations

Revisiting R^2

- R^2 is the proportion of variability in y explained by the model:
  - large: linear relationship between x and y exists
  - small: evidence provided by the data may not be convincing

DATA ANALYSIS & STATISTICAL INFERENCE Duke

(7) Unit 6 part 4 (2)

- Two ways to calculate R^2:

(1) Using correlation = square of correlation coefficient

(2) From the definition: $\frac{\text{explained variability}}{\text{total variability}}$

Variability partitioning - linear regression

Partitioning the variability in y to explained and unexplained variability requires ANOVA

Example. ANOVA output:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bio IQ	1	5231.13	5231.13	87.56	0.00000
Residuals	25	1493.53	59.74		
Total	26	6724.66			

Sum of squares

Total variability in y: $SS_{Tot} = \sum (y - \bar{y})^2 = 6724.66$

Unexplained variability in y (residuals):

$SS_{Res} = \sum (y - \hat{y})^2 = \sum e_i^2 = 1493.53$

Explained variability in y: $SS_{Reg} = 6724.66 - 1493.53 = 5231.13$

Degrees of freedom:

Total degrees of freedom: $df_{Tot} = 27 - 1 = 26$

(72) Unit 6 Part 4 (2)

Regression degree of freedom: $df_{Reg} = 1$ (Only 1 predictor)

Residual degree of freedom: $df = 26 - 1 = 25$

Mean Squares

MS regression: $MS_{Reg} = \frac{SS_{Reg}}{df_{Reg}} = \frac{5231.13}{1} = 5231.13$

MS residual: $MS_{Res} = \frac{SS_{Res}}{df_{Res}} = \frac{1493.53}{25} = 59.74$

F statistic: ratio of explained to unexplained variability

$$F_{(1,25)} = \frac{MS_{Reg}}{MS_{Res}} = 87.56$$

Hypothesis test

$H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ small p-value = reject H_0

The data provide convincing evidence that the slope is significantly different than 0, i.e. the explanatory variable is a significant predictor of the response variable.

R^2: $R = 0.882$ so $R^2 \simeq 0.78$

78% of the variability in foster twins IQs can be explained by the model, i.e. by the bio twins IQs.

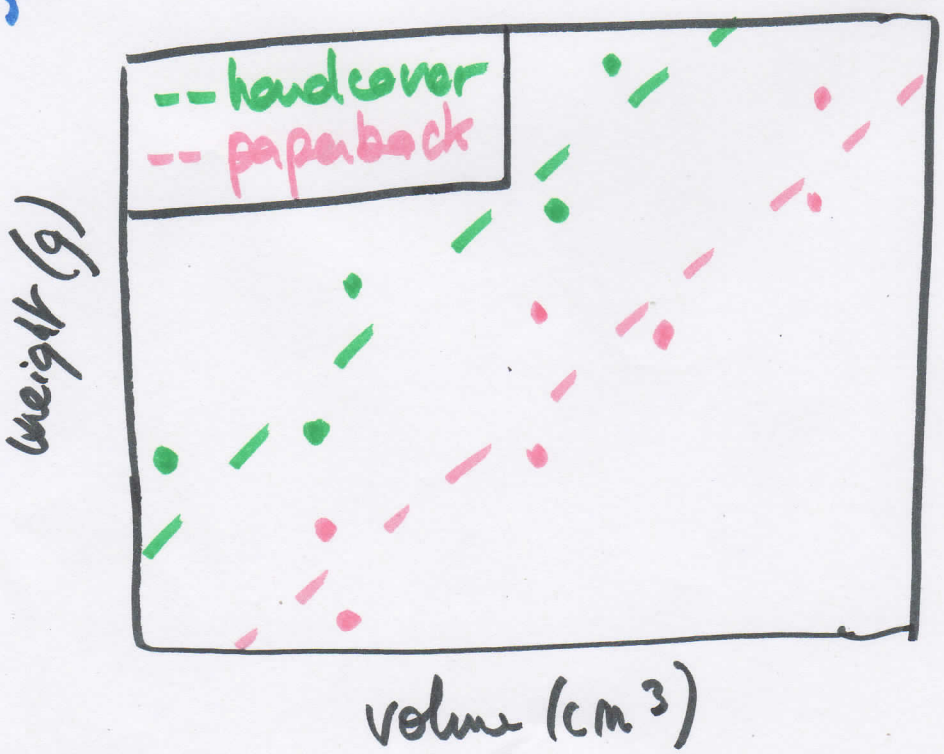
Multiple predictors - Example

Predict book weight from volume and cover (pb or hb)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	197.96	59.19	3.34	0.01
volume	0.72	0.06	11.67	0.00
cover: pb	-184.05	40.49	-4.55	0.00

$\widehat{weight} = 197.96 + 0.72\ volume - 184.05\ cover:pb$

- For hardcover books (ref level): plug in 0 for cover
  $\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 0$
- For paperback books: plug in 1 for cover
  $\widehat{weight} = 197.96 + 0.72\ volume - 184.05 \times 1$



DATA ANALYSIS & STATISTICAL INFERENCE

Duke

(74) Unit 7 Part 1 (1) (2)

Interpreting the regression parameter: slope - Example

Slope of volume: All else held constant, for each $1 cm^3$ increase in volume the model predicts the book's weight to be heavier on average by 0.72 grams.

Slope of cover: All else held constant, the model predicts that paperback books weigh 184.05 grams lower than hardcover books, on average.

Interpreting the regression parameters: intercept. Example

Intercept: Hardcover books with no volume are expected on average to weigh 198 grams.

- Meaningless in context, serves to adjust the height of the line.

Adjusted R^2

$$R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1}\right)$$

k = number of predictors

R^2 vs. adjusted R^2

- When any variable is added to the model, R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

(75) Unit 7 Part 1(2)

Properties of adjusted R^2

- k is never negative, then adjusted $R^2 < R^2$
- Adjusted R^2 applies a penalty for the number of predictors included in the model
- We choose models with higher adjusted R^2 over other

Adjusted R^2 - Example

Get ANOVA output in R: anova(linear model)

Calculate adjusted R^2 for the multiple linear model predicting % living in poverty from % female householders and % white. n = 51

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female-house	1	132.57	132.57	18.74	0.00
white	1	8.21	8.21	1.16	0.29
Residuals	48	339.47	7.07		
Total	50	480.25			

$$R^2_{adj} = 1 - \left(\frac{SSE}{SST} \times \frac{n-1}{n-k-1}\right)$$

$$= 1 - \left(\frac{339.47}{480.25} \times \frac{51-1}{51-2-1}\right) = 0.26$$

DATA ANALYSIS & STATISTICAL INFERENCE

Duke

(46) Unit 7 Part 1(3) Part 2(1)

Collinearity

- Two predictor variables are said to be collinear when they are correlated with each other.
- Predictors are also called independent variables, so they should be independent of each other.
- Inclusion of collinear predictors (also called multicollinearity) complicates model estimation.

Parsimony

- Avoid adding predictors associated with each other because often times the addition of such variable brings nothing new to the table.
- Prefer the simplest best model, ie. the parsimonious model.
  - Occam's razor: Among competing hypotheses, the one with the fewest assumptions should be selected
- Addition of collinear variables can result in bloated estimates of the regression parameters
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to control for correlated predictors.

Inference for MLR (Multilinear regression)

t-statistic for the slope: $T = \frac{b_1 - 0}{SE_{b_1}}$ $df = n - k - 1$

k = nb of predictors

(77) Unit 7 Part 2(1)

Inference for the multilinear regression model as a whole

$H_0: \beta_1 = \beta_1 = \ldots = \beta_k = 0$ (k slopes)

H_A: At least one β_i is different than 0

F-statistics: 29.74 on 4 and 429 DF, p value: < 0.02..

Since p-value < 0.05, the model as a whole is significant

- The F test yielding a significant result doesn't mean the model fits the data well, it just means at least one of the β_i is non-zero
- The F-test not yielding a significant result doesn't mean individuals variables included in the model are not good predictors of y, it just means that the combination of these variables doesn't yield a good model.

Hypothesis testing for slopes (MLR)

$H_0: \beta_1 = 0$, when all other variables are included in the model

$H_A: \beta_1 \neq 0$ → calculate T, df and then p-value

$T = 2201$ $df = 429$ In R: pt(2201, df = 429, lower.tail = FALSE) * 2

Confidence intervals for slopes

point estimate ± margin of error

$b_1 \pm t^*_{df} SE_{b_1}$

Example in R

df = 429. 95% CI

qt(0.025, df = 429)

(78) Unit 7 Part 2 (1) Part 3 (1)

Example: interpret the 95% CI for the slope of "mom_work"

CI: (-2.09, 7.17)

We are 95% confident that, all else being equal, the model predicts that children whose moms worked during the first three years of their lives score 2.09 pts lower to 7.77 pts higher than those whose moms did not work

Stepwise model selection (MCR)

backwards elimination: start with a full model (containing all predictors), drop one predictor at a time until the parsimonious model is reached

forward selection: start with an empty model and add one predictor at a time until the parsimonious model is reached

- Criteria:
  - p-value, adjusted R^2 (what we'll use)
  - AIC, BIC, DIC, Bayes Factor, Mallow's C_p (beyond our scope)

Backward elimination - adjusted R^2

- Start with the full model
- Drop one variable at a time and record adjusted R^2 of each smaller model
- Pick the model with the highest increase in adjusted R^2

Unit 7 Part 3 (1)

- Repeat until none of the models yield an increase in adjusted R^2

Backwards elimination - p-value

- Start with the full model
- Drop the variable with the highest p-value and refit a smaller model
- Repeat until all variables left in the model are significant

⚠ If you have a categorical variable with multiple levels, you cannot drop part of the levels. You have to either keep the variable as a whole or drop it as a whole.

Forward selection - adjusted R^2

- Start with single predictor regressions of response vs. each explanatory variable
- Pick the model with the highest adjusted R^2
- Add the remaining variables one at a time to the existing model, and pick the model with the highest adjusted R^2
- Repeat until the addition of any of the remaining variables does not result in a higher adjusted R^2

Forward selection - p-value

- Start with single predictor regressions of response vs. each explanatory variable

80 Unit 7 Part 3(1)

- Pick the variable with the lowest significant p-value
- Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
- Repeat until any of the remaining variables do not have a significant p-value.

Adjusted R^2 vs. p-value

p-value: significant predictors

Adjusted R^2: more reliable predictions

- p-value method depends on the (somewhat arbitrary) 5% significance level cutoff
  - different significance level → different model
  - Used commonly since it requires fitting fewer models (in the more commonly used backwards selection approach)

Expert opinion

- Variables can be included in (or eliminated from) the model based on expert opinion
- If you are studying a certain variable, you might choose to leave it in the model regardless of whether it's significant or yield a higher adjusted R^2.

(81) Unit 7 Part 4(1)

Diagnostics for MLR

cannot expect linearity from cat. var.

(1) Linear relationships between (numerical) x and y

- Each (numerical) explanatory variable linearly related to the response variable
- Check using residual plots (e vs. x)
  - Looking for a random scatter around 0
  - Instead of scatterplot of y vs. x: allows for considering the other variable that are also in the model, and not just the bivariate relationship between a given x and x
  - with all variable needed

K: plot(model$residuals ~ dataset$variable1)

(2) Nearly normal residuals with mean 0

- Some residuals will be positive and some negative
- On a residuals plot we look for random scatter of residuals around 0
- This translates to a nearly normal distribution of residuals centered at 0
- Check using histogram or normal prob. plot

K: hist(model$residuals) → Histogram of residuals

qqnorm(model$residuals)
qqline(model$residuals)
] normal probability plot of residuals

(3) Constant variability of residuals

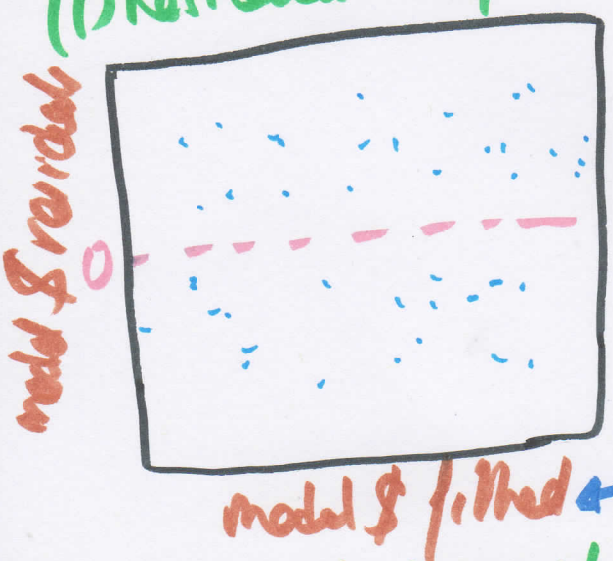
- Residuals should be equally variable for low and high values of the predicted response variable

(82) Unit 7 Part 4 (1)

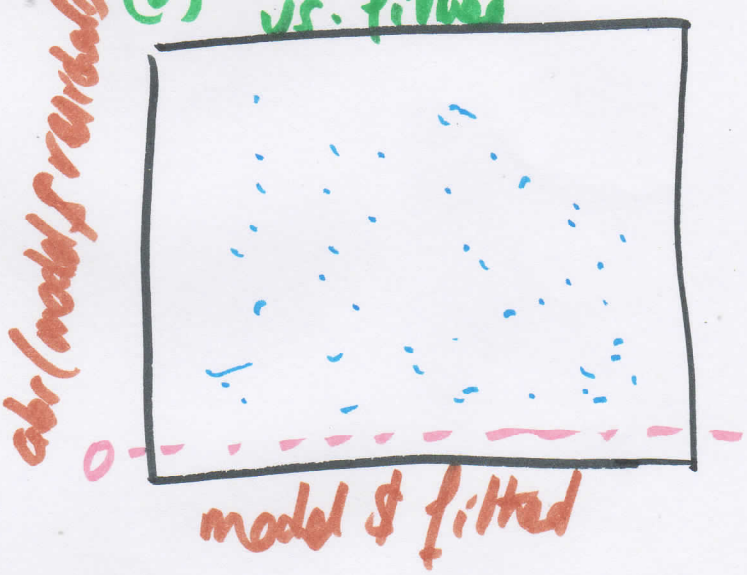
- Check using residuals plots of residuals vs. predicted (or $\hat{y}$)
  - Residuals vs. predicted instead of residuals vs. x because it allows for considering the entire model (with all explanatory variables) at once
  - Residuals randomly scattered in a band with a constant width around 0 (no fan shape)
  - Also worthwhile to view absolute value of residuals vs. predicted to identify unusual observations easily
    - a triangle shape = no

R: (1) plot(model$residuals ~ model$fitted)
(2) plot(abs(model$residuals) ~ model$fitted)

(1) Residuals vs. fitted



(2) Absolute value of residuals vs. fitted



(4) Independent residuals

- Independent residuals → independent observations
- If time series structure is suspected check using residuals vs. order of data collection
- If not, think about how the data are sampled

R: plot(model$residuals)

Review: frequentist vs. bayesian inference

Population: M&Ms. Percentage of yellow M&Ms is either 10% or 20%.

Decision table	True state of the population	
Decision	% yellow = 10%	% yellow = 20%
% yellow = 10%	☺	☹
% yellow = 20%	☹	☺

① Frequentist inference

Hypotheses: H_0: 10% yellow M&Ms
H_A: 20% yellow M&Ms

Test statistic: number of yellow M&Ms observed in the sample

p-value: probability of observing this many or more yellow M&Ms given that the null hypothesis is true.

$\alpha = 0.05$

p-value (for $n = 5$):
$P(1 \text{ or more yellows} \mid n = 5, p = 10\%) \approx 0.26$

R: sum(dbinom(1:5, 5, 0.1))

nb of yellow in sample	P(k or more | 10% yellow)
n = 5 : 1	0.41
n = 10 : 2	0.26
n = 15 : 3	0.18
n = 20 : 4	0.13

DATA ANALYSIS & STATISTICAL INFERENCE

Date

(8/4) Unit 8

② Bayesian inference

- Start over with the 1:1 odds that the percentage of yellows is 10% : 20% (the prior probabilities)

$P(10\%.\ \text{yellow}) = 0.5$

$P(20\%.\ \text{yellow}) = 0.5$

- Use the same data and Bayes' theorem to calculate the probability of either of the hypotheses being true given the observed data (the posterior probabilities)

Bayes' theorem: $P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

$$P(10\%.\ \text{yellow} \mid \text{data}) = \frac{P(\text{data} \wedge 10\%.\ \text{yellow})}{P(\text{data})}$$

$$= \frac{P(\text{data} \mid 10\%.\ \text{yellow}) \times P(10\%.\ \text{yellow})}{P(\text{data})}$$

$$P(20\%\ \text{yellow} \mid \text{data}) = \frac{P(\text{data} \mid 20\%.\ \text{yellow}) \times P(20\%.\ \text{yellow})}{P(\text{data})}$$

$$= 1 - P(10\%.\ \text{yellow} \mid \text{data})$$

For $n = 5$

prior: $P(p = 10\%) = 0.5$; $P(p = 20\%) = 0.5$

posterior: $P(p = 10\% \mid \text{data})$

In R: dbinom(1, 5, 0.1) $\simeq 0.33$

DATA ANALYSIS & STATISTICAL INFERENCE Duke

(85) Unit 8

$$P(p = 10\% \mid \text{data}) = \frac{P(\text{data} \mid p = 10\%) \times P(p = 10\%)}{P(\text{data} \mid p = 10\%) \times P(p = 10\%) + P(\text{data} \mid p = 20\%) \times P(p = 20\%)}$$

$$= \frac{0.33 \times 0.5}{0.33 \times 0.5 + 0.41 \times 0.5} \geq 0.44$$

Summary results = Frequentist vs. Bayesian

	FREQUENTIST	BAYESIAN	
Nb of yellow MxMr in sample	P(k or more | 10% yellow)	P(10% yellow | n, k)	P(20% yellow | n, k)
n = 5 : 1	0.41	0.44	0.56
n = 10 : 2	0.26	0.39	0.61
n = 15 : 3	0.18	0.34	0.66
n = 20 : 4	0.13	0.29	0.71

F: choose Ho p-value > 0.05

B: choose H_A $P(20\% \text{ yellow} \mid n, k) > P(10\% \text{ yellow} \mid n, k)$

⚠ Results diverge

The true result here is 20% yellow.